technische universität
dortmund

Intuitive Joint Priors for High-Dimensional Bayesian Linear Models

Dissertation zur Erlangung des Doktorgrades Dr. rer. nat.
der Fakultät Statistik
der Technischen Universität Dortmund

Vorgelegt von

Javier Enrique Aguilar Romero

Vorsitzender der Prüfungskommission:

Prof. Dr. Jörg Rahnenführer

Prüfende der Dissertation:

1. Prof. Dr. Paul-Christian Bürkner
2. Prof. Dr. Katja Ickstadt
3. Prof. Dr. Aki Vehtari

Datum der mündlichen Prüfung: 25.03.2026

Wer immer strebend sich bemüht,
Den können wir erlösen.

—Johann Wolfgang von Goethe, *Faust II*

Abstract

In high-dimensional regression, where the number of predictors is large relative to the sample size, the prior distribution can have a decisive effect on posterior inference, making careful prior specification essential. In many applications, the regression is also structured, for example through multilevel terms, grouped predictors, interactions, or covariates with substantial dependence. Standard shrinkage priors are typically formulated for exchangeable coefficients and do not provide a direct mechanism to encode such structure or to induce joint regularization across model terms. This motivates priors that act on structured components rather than on coefficients in isolation, so that regularization is applied both across and within the components defining the linear predictor. Moreover, hyperparameters are often specified at the coefficient level, which can obscure implications for predictive quantities and complicate calibration against domain knowledge. These considerations call for prior specification strategies that allow one to reason directly about predictive behavior through interpretable quantities, and that translate such beliefs into coherent joint priors on the regression parameters.

This cumulative thesis develops continuous global-local shrinkage priors for high-dimensional Bayesian linear models with structured predictors through a predictive-first framework. We place priors on the coefficient of determination R^2 and translate these specifications to model parameters via variance decomposition mechanisms. The resulting priors induce joint regularization across structured components while retaining properties required for adaptive regularization in high dimensions.

Across four papers, we develop a family of priors that extends R^2-based variance decompositions to multilevel, grouped, and dependent-design settings. We introduce the R2D2M2 prior for multilevel models, which jointly regularizes population-level and varying effects and supports settings in which both sets of effects are high-dimensional. We propose generalized decomposition priors that explicitly model dependence among the variance components of the linear predictor, expanding the range of joint shrinkage behavior that can be expressed through the decomposition. For grouped predictors, we develop a two-stage allocation that separates between-group sparsity from within-group heterogeneity. We further generalize shrinkage priors to incorporate dependence at the level of covariates by encoding correlation structure, together with data-driven strategies for specifying this structure in practice.

We study theoretical properties of the proposed priors, including their implied shrinkage and concentration behavior, and show that they place substantial mass near the origin while retaining heavy tails to accommodate large signals. We further characterize shrinkage and its relationship to hyperparameters via shrinkage factors, and we generalize the notion of effective model size to the structured settings considered. Simulation studies and applied analyses

investigate when, and to what extent, incorporating structure improves parameter recovery and out-of-sample predictive performance.

Overall, this thesis provides a principled framework for prior specification in structured high-dimensional regression, enabling practitioners to express assumptions through interpretable predictive quantities while maintaining the regularization behavior needed for reliable inference.

Zusammenfassung

In der hochdimensionalen Regression, in der die Anzahl der Prädiktoren groß im Verhältnis zur Stichprobengröße ist, kann die Priorverteilung einen maßgeblichen Einfluss auf die posterioren Schlussfolgerungen haben, sodass eine sorgfältige Prior-Spezifikation entscheidend ist. In vielen Anwendungen ist die Regression zudem strukturiert, etwa durch Mehrebenenterme, gruppierte Prädiktoren, Interaktionen oder Kovariaten mit ausgeprägter Abhängigkeit. Standardmäßige Shrinkage-Priors werden typischerweise für austauschbare Koeffizienten formuliert und bieten keinen direkten Mechanismus, um eine solche Struktur zu kodieren oder eine gemeinsame Regularisierung über Modellterme hinweg zu induzieren. Dies motiviert Priors, die auf strukturierten Komponenten wirken, statt auf Koeffizienten in Isolation, sodass Regularisierung sowohl zwischen als auch innerhalb der Komponenten erfolgt, die den linearen Prädiktor definieren. Zudem werden Hyperparameter häufig auf Koeffizientenebene spezifiziert, was die Konsequenzen für prädiktive Größen verschleiern und die Kalibrierung anhand von Domänenwissen erschweren kann. Diese Überlegungen erfordern Strategien der Prior-Spezifikation, die es erlauben, direkt über das prädiktive Verhalten anhand interpretierbarer Größen nachzudenken und solche Annahmen in kohärente gemeinsame Priors auf die Regressionsparameter zu übersetzen.

Diese kumulative Dissertation entwickelt kontinuierliche global-lokale Shrinkage-Priors für hochdimensionale bayessche lineare Modelle mit strukturierten Prädiktoren im Rahmen eines prädiktiv-zentrierten Ansatzes. Wir setzen Priors auf das Bestimmtheitsmaß R^2 und übertragen diese Spezifikationen mittels Varianzzerlegungsmechanismen auf die Modellparameter. Die resultierenden Priors induzieren eine gemeinsame Regularisierung über strukturierte Komponenten hinweg und bewahren zugleich Eigenschaften, die für adaptive Regularisierung in hohen Dimensionen erforderlich sind.

Über vier Arbeiten hinweg entwickeln wir eine Familie von Priors, die R^2-basierte Varianzzerlegungen auf Mehrebenen-, Gruppen- und abhängige Design-Situationen erweitert. Wir führen den R2D2M2-Prior für Mehrebenenmodelle ein, der Populationseffekte und zufällige Effekte gemeinsam regularisiert und Settings unterstützt, in denen beide Effektmengen hochdimensional sind. Wir schlagen verallgemeinerte Zerlegungspriors vor, die Abhängigkeiten zwischen den Varianzkomponenten des linearen Prädiktors explizit modellieren und damit das Spektrum gemeinsamer Shrinkage-Verhaltensweisen erweitern, die über die Zerlegung ausgedrückt werden können. Für gruppierte Prädiktoren entwickeln wir eine zweistufige Allokation, die zwischen Gruppen-Sparsität von innerhalb der Gruppe liegender Heterogenität trennt. Darüber hinaus verallgemeinern wir Shrinkage-Priors, um Abhängigkeiten auf Kovariatenebene durch Kodierung von Korrelationsstruktur zu berücksichtigen, zusammen mit datengetriebenen Strategien zur Spezifikation dieser Struktur in der Praxis.

Wir untersuchen theoretische Eigenschaften der vorgeschlagenen Priors, einschließlich ihres impliziten Shrinkage- und Konzentrationsverhaltens, und zeigen, dass sie substanzielle Masse nahe dem Ursprung platzieren, während sie gleichzeitig schwere Tails beibehalten, um große Signale zuzulassen. Ferner charakterisieren wir Shrinkage und dessen Zusammenhang mit Hyperparametern über Shrinkage-Faktoren und verallgemeinern den Begriff der effektiven Modellgröße auf die betrachteten strukturierten Settings. Simulationsstudien und angewandte Analysen untersuchen, wann und in welchem Ausmaß die Berücksichtigung von Struktur die Parameterschätzung und die Vorhersagegüte außerhalb der Stichprobe verbessert.

Insgesamt liefert diese Dissertation einen prinzipiellen Rahmen für die Prior-Spezifikation in strukturierter hochdimensionaler Regression, der es Anwenderinnen und Anwendern ermöglicht, Annahmen über interpretierbare prädiktive Größen auszudrücken und zugleich das Regularisierungsverhalten zu bewahren, das für verlässliche Inferenz erforderlich ist.

Acknowledgements

First, I would like to thank my supervisor, Paul Bürkner, for all his support, guidance, patience and empathy throughout these years. You provided me with an environment where my curiosity could flourish, where I could openly learn and gradually forge my identity as a scientist. Thank you for believing in me and my ideas.

I gratefully acknowledge financial support from the Deutsche Forschungsgemeinschaft (DFG, Project-500663361), the Department of Statistics at TU Dortmund, ELLIS through the ELISE mobility program, the Cluster of Excellence SimTech, and travel grants from the Institute of Mathematical Statistics and the International Society for Bayesian Analysis.

I am grateful to my dissertation committee for their time and feedback. I thank Jörg Rahnenführer for chairing the committee.

I am thankful to Aki Vehtari and David Kohns for their collaboration. Aki, thank you for your guidance, for sharing your ideas with me, and for your encouragement. David, thank you for the stimulating discussions and for your openness to exploring new research ideas together.

I would like to thank Katja Ickstadt for her guidance on writing this thesis, for her kind spirit, and for helping me navigate my next steps. I owe special thanks to Fabrizio Ruggeri, Andrés Christen, and Leticia Ramírez, whose encouragement set me on this path. You showed me what kind of academic I wanted to become. I am especially grateful to Fabrizio and the IMU for supporting scholars from underrepresented communities worldwide.

I thank my colleagues for sharing their time and perspectives, for showing me how others navigate academic life, and for the afternoon coffees. To my friends Carbajal, Fabi, Harold, Martha, Rafa, Ricardo, Víctor, thank you for being there for me.

A mi familia, que es el principio y el fin de las cosas. A mis padres, Javier y Mirian, porque el amor que me han brindado es la verdadera semilla de cada mérito en mí. Gracias por abrazarme. A mis hermanos, Dennis y Elías, y a mi hermana, Nadia Jamil, que son el espejo de mi alma y con quienes puedo reír y ser yo mismo. Son todo mi corazón.

A María Fernanda, my rainbow, how good it is to know you're like me. Siempre has sido el hogar y descanso que mi alma requería. A quien fui, he sido y seré, gracias por ser valiente siempre. A Dios, por darme paz y sostenerme en sus manos siempre.

Whoever you are,
the world offers itself to your imagination
over and over announcing your place
in the family of things. —Mary Oliver

List of Articles

This cumulative thesis is based on the following four manuscripts:

Article 1: Aguilar, J. E., and Bürkner, P.-C. (2023). Intuitive joint priors for Bayesian linear multilevel models: the R2D2M2 prior. *Electronic Journal of Statistics* 17(1): 1711–1767. DOI: https://doi.org/10.1214/23-EJS2136. Aguilar and Bürkner (2023).

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Author contributions: Javier Enrique Aguilar developed the methodology, derived the theoretical results, designed and conducted the simulation studies, interpreted the results, and wrote the manuscript. Paul-Christian Bürkner proposed the initial idea, supervised the project, contributed to the design and interpretation of the simulation study, and reviewed, edited, and approved the final manuscript.

Article 2: Aguilar, J. E., and Bürkner, P.-C. (2025). Generalized decomposition priors on R^2. *Bayesian Analysis*. DOI: https://doi.org/10.1214/25-BA1524. Aguilar and Bürkner (2025b).

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Author contributions: Javier Enrique Aguilar and Paul-Christian Bürkner jointly conceived the initial idea. Javier Enrique Aguilar developed the methodology, derived the theoretical results, designed, implemented, and conducted the simulation studies, interpreted the results, and wrote the manuscript. Paul-Christian Bürkner supervised the project, contributed to the design of the simulation study, and reviewed, edited, and approved the final manuscript.

Article 3: Aguilar, J. E., Kohns, D., Vehtari, A., and Bürkner, P.-C. (2025). R2 priors for grouped variance decomposition in high-dimensional regression. Preprint. Manuscript submitted for publication. Available at https://arxiv.org/abs/2507.11833. Aguilar et al. (2025).

Author contributions: Javier Enrique Aguilar developed the methodology, derived the theoretical results and proofs, designed and implemented the simulation studies, analyzed and interpreted the results, and wrote the manuscript. David Kohns conceived the initial idea, developed the methodology, assisted with simulation

design and implementation, co-wrote parts of the manuscript, and reviewed, edited, and approved the final version. Aki Vehtari provided advisory feedback, reviewed the manuscript, and reviewed, edited, and approved the final version. Paul-Christian Bürkner reviewed and approved the final manuscript.

Article 4: Aguilar, J. E., and Bürkner, P.-C. (2025). Dependency-aware shrinkage priors for high-dimensional regression. Preprint. Manuscript submitted for publication. Available at https://arxiv.org/abs/2505.10715. Aguilar and Bürkner (2025a).

Author contributions: Javier Enrique Aguilar and Paul-Christian Bürkner jointly conceived the initial idea. Javier Enrique Aguilar developed the methodology, derived the theoretical results, designed, implemented, and conducted the simulation studies, interpreted the results, and wrote the manuscript. Paul-Christian Bürkner supervised the project, contributed to the design of the simulation study, and reviewed, edited, and approved the final manuscript.

Below, I list additional articles completed during my doctoral studies that are not included in this cumulative thesis.

Article 5: Reiser, P., Aguilar, J. E., Guthke, A., and Bürkner, P.-C. (2025). Uncertainty quantification and propagation in surrogate-based Bayesian inference. *Statistics and Computing*. DOI: https://doi.org/10.1007/s11222-025-10597-8 . Reiser et al. (2025).

Article 6: Matabuena, M., Ghosal, R., Aguilar, J. E., Wagner, R., Merino, C. F., Castro, J. S., Zipunnikov, V., Onnela, J. P., and Gude, F. (2025). Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics. *Scientific Reports*. DOI: https://doi.org/10.1038/s41598-025-18119-2 . Matabuena et al. (2025).

Article 7: Fazio, L., Scholz, M., Aguilar, J. E., and Bürkner, P.-C. (2025). Primed priors for simulation-based validation of Bayesian models. Preprint. Manuscript submitted for publication. Available at https://arxiv.org/abs/2408.06504 . Fazio et al. (2025).

Article 8: Mukherjee, S., Aguilar, J. E., Zago, M., Classen, M., and Bürkner, P.-C. (2025). Latent variable estimation with composite Hilbert space Gaussian processes. Preprint. Manuscript submitted for publication. Available at https://arxiv.org/abs/2510.25371 . Mukherjee et al. (2025).

Contents

Abstract v

Zusammenfassung vii

Acknowledgements ix

List of Articles xi

1 Introduction 1

2 Background 9
2.1 Bayesian inference . . . 9
2.1.1 Predictive distributions . . . 10
2.1.1.1 The prior predictive distribution . . . 10
2.1.1.2 The posterior predictive distribution . . . 13
2.2 High-dimensional linear regression . . . 14
2.3 Shrinkage priors . . . 17
2.3.1 Global-local formulations . . . 17
2.3.1.1 Marginal distributions . . . 20
2.3.1.2 Tweedie's formula and posterior means . . . 22
2.3.2 Priors on the intercept and variance . . . 23
2.3.3 Implied priors on lower-dimensional quantities . . . 23
2.3.4 Hyperparameter specification . . . 24
2.3.5 Shrinkage factors . . . 25
2.3.6 Effective number of nonzero coefficients . . . 26
2.3.7 Variable selection . . . 27
2.3.8 Computational methods . . . 28
2.4 Priors over R^2 . . . 29
2.5 The R2D2 prior . . . 30

3 Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior 35

4 Generalized decomposition priors on R^2 41

5 R^2 priors for grouped variance decomposition in high-dimensional regression 45

6 Dependency-aware shrinkage priors for high-dimensional regression 49

7 Discussion 53

List of Figures 63

Bibliography **65**

Publications **83**

Article I: Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior 85

Article II: Generalized decomposition priors on R^2 145

Article III: R2 priors for grouped variance decomposition in high-dimensional regression 201

Article IV: Dependency-Aware shrinkage priors for high-dimensional regression. . . 279

Affidavits **327**

1 Introduction

Linear regression is arguably the most widely used statistical model across scientific disciplines (Gelman and Hill, 2006; Gelman et al., 2013). Their ubiquity stems not from simplicity but from genuine versatility. Many complex relationships are well approximated locally by linear functions, nonlinear models often reduce to linear forms after transformation, and modern extensions such as splines and basis expansions allow linear frameworks to represent rich functional forms (Wasserman, 2004). The continued centrality of linear regression in applied statistics reflects not conservatism but the model's proven ability to deliver interpretable, reliable inference across a wide range of applications (Harrell, 2013; Gelman et al., 2020a).

When the number of predictors p is large relative to the sample size n, however, standard estimation procedures break down. The design matrix becomes rank-deficient or ill-conditioned and the model is not identifiable from the likelihood alone. Yet high-dimensional regression is not a pathological edge case but increasingly the norm, mainly because modern data collection produces many candidate predictors while sample sizes remain limited. In genomics, studies routinely measure tens of thousands of gene expression levels (Baek et al., 2026). Similar regimes appear in neuroimaging, where analyses can involve thousands of voxel-level measurements (Soares et al., 2016), and in text analysis, where bag-of-words representations routinely lead to feature spaces with thousands of dimensions (Joachims, 1998). Spatial and spatio-temporal applications also become high dimensional once fine-resolution latent effects or basis representations are introduced (Lindgren et al., 2011; Lindgren and Rue, 2015). Across these areas it is common that the number of candidate effects p is comparable to, or exceeds, the number of observations n (Buehlmann et al., 2014).

The Bayesian response to this dimensionality challenge is to encode prior beliefs about the structure of regression coefficients, providing additional information that makes inference well-defined even when the likelihood alone cannot identify parameters uniquely. As statisticians, we would like our models to work reliably in these complex high-dimensional settings, delivering stable estimates and transparent uncertainty quantification. The question is: what prior distributions make this possible?

A guiding principle that has proven remarkably successful is sparsity. The idea is that even when the dimension p is very large, only a relatively small subset of predictors carries genuine signal (Buehlmann et al., 2014; Wainwright, 2019). This assumption is convenient and often substantively plausible, and it has proven successful in many applications. For instance, in genome-wide association studies, only a small number of genetic variants among thousands may be causally related to a phenotype (Malone et al., 2011).

The sparsity principle has motivated a large literature on variable selection and regularization (Giraud, 2014; Buehlmann et al., 2014). From a Bayesian perspective, sparsity is encoded through priors that place substantial mass near zero for most coefficients, while still allowing a minority of coefficients to escape shrinkage (Tadesse and Vannucci, 2021). Continuous global-local shrinkage priors have become a widely used way to achieve this behavior through scale-mixture constructions (Andrews and Mallows, 1974; West, 1987; Pas, 2021). In these constructions, each coefficient receives a local scale parameter that determines how strongly it is regularized, while a global scale parameter shared across all coefficients controls the overall strength of shrinkage (Polson and Scott, 2011; Griffin and Brown, 2013). Well established examples include the Horseshoe prior (Carvalho et al., 2009), the Normal-Gamma prior (Brown and Griffin, 2010), beta-mixtures of Gaussians (Armagan et al., 2011), the Dirichlet-Laplace prior (Bhattacharya et al., 2015), the Regularized Horseshoe (Piironen and Vehtari, 2017), the Beta-Prime prior (Bai and Ghosh, 2019), and the R2D2 prior (Zhang et al., 2020). These priors differ in their specific distributional choices but share this core global-local structure.

These priors have proven successful along multiple dimensions. Empirically, they deliver strong predictive performance and meaningful parameter recovery across diverse applications (Boss et al., 2023). Theoretically, they satisfy important properties including posterior consistency and near-minimax contraction rates under appropriate model and sparsity assumptions (Ghosh, 2003; Pas et al., 2014, 2016; Ghosh and Chakrabarti, 2017). Computationally, they admit efficient implementations, such as conjugate blocked Gibbs sampling (Bhattacharya and Johndrow, 2021). They also work well with modern Hamiltonian Monte Carlo methods that enable automatic, gradient-based posterior exploration (Neal, 2011; Hoffman and Gelman, 2014). This combination of empirical performance, theoretical grounding, and computational tractability has made continuous shrinkage priors the method of choice for high-dimensional Bayesian regression.

Their increasing adoption in applied Bayesian analysis provides further evidence of this success. Implementations in widely used probabilistic programming languages, including Stan (Stan Development Team, 2026), PyMC (Abril-Pla et al., 2023), and Turing (Ge et al., 2018), make these priors accessible to a broad set of users and integrate them into routine modelling workflows.

Development of priors for high-dimensional regression has largely focused on separating signal from noise when the number of candidate predictors is large (Castillo and Vaart, 2012). The aim is adaptive regularization that shrinks irrelevant coefficients strongly while preserving meaningful effects, without requiring users to specify in advance which predictors matter (Banerjee et al., 2021). A key insight is that adaptation to sparsity can be learned from the data when the prior is structured appropriately. Theoretical and empirical work shows that heavy-tailed local scales can protect large signals, while a global scale with substantial mass near zero provides strong shrinkage of noise, which together support reliable recovery in high-

dimensional settings (Pas, 2021; Castillo, 2024). This understanding emerged through careful analysis of how distributional choices for the scales translate into marginal concentration near zero, tail behavior, and posterior contraction rates.

Yet among the many priors proposed, relatively few begin by specifying priors on predictive quantities that practitioners can interpret directly and then propagate this uncertainty to the regression coefficients. This predictive-first perspective builds on the mature understanding of coefficient-level shrinkage priors, in particular which concentration and tail properties are essential for reliable high-dimensional inference (Zhang et al., 2020; Manderson and Goudie, 2026).

Specifying priors on predictive quantities is appealing because it matches how applied questions are posed and because it connects directly to prior predictive distributions. Prior predictive checks provide a concrete way to assess what a prior implies before seeing the data and to evaluate whether induced predictions are compatible with domain knowledge (Gelman et al., 2017, 2020b; Mikkola et al., 2021; Linero, 2023). This is often easier than reasoning directly about coefficient-level hyperparameters, whose implications can be opaque even in moderate dimensions and become harder to anticipate as the number of coefficients increases (Efron, 1978; Bhadra et al., 2016).

However, placing priors on quantities of interest does not by itself solve the high-dimensional inference problem. Two fundamental challenges must be addressed. First, we need a principled mechanism for propagating uncertainty from predictive quantities to model parameters, ideally inducing a joint distribution on the coefficient vector. Second, and more subtly, nothing guarantees that a prior designed for interpretability will retain the theoretical properties essential for high-dimensional inference. A prior can be easy to specify and communicate while still failing to adequately shrink noise or detect signals in high dimensions (Efron, 1978; Gelman, 2006).

In this thesis we pursue this idea through priors on the coefficient of determination R^2. The quantity R^2 is widely used as a summary of model fit and has an immediate interpretation in terms of explained variability. Placing a prior on R^2 therefore provides a clean and communicable way to express beliefs about signal strength, and it yields prior predictive implications that can be inspected and calibrated. The central goal is to use this predictive-level specification as the starting point and then propagate it to a joint prior on regression coefficients that remains suitable for high-dimensional inference. Our focus is to develop methods for structured high-dimensional regression, including multilevel models, grouped predictors, and settings with predictor dependencies.

The difficulty of prior specification becomes sharper once the regression problem has structure. Structure changes what coefficients represent, the objective regularization should pursue, and how hyperparameters translate to prior predictive distributions. As the complexity

of applied regression problems has grown, structure has become the rule rather than the exception (Griffin and Brown, 2017, 2021; Boss et al., 2023; Griffin and Hoff, 2024). Standard shrinkage priors were not designed with these structures in mind. Despite their success in simpler settings, they treat all coefficients symmetrically and do not account for mechanisms that reflect groupings, hierarchical organizations, or forms of dependence among coefficients. Extensions to structured settings exist, particularly in the context of groups, but they often sacrifice theoretical guarantees, interpretability of hyperparameters, or the ability to represent a sensible baseline. Moreover, few provide the joint regularization across multiple components that structured problems require (Fuglstad et al., 2020).

Multilevel models move beyond a single coefficient per predictor and allow effects to vary across groups (Gelman and Hill, 2006). This family of models is now standard in fields ranging from education research to clinical trials, driven both by the hierarchical nature of data collection (such as students within schools, patients within hospitals) and by the desire to learn both population-level and group-specific effects (Harrell, 2013; Gelman et al., 2020a). If varying effects are introduced broadly, for example by allowing group-specific deviations for many predictors, the parameter dimension grows quickly with the number of groups (Lin et al., 2020). At the same time, many groups may be small. This creates a setting where parts of the model are weakly identified by the data and where the prior plays a central role in determining which variation is treated as signal and which is treated as noise. Moreover, prior specification in multilevel models has traditionally treated different terms independently: even though varying effects within each term receive a joint prior (typically multivariate normal), the priors across terms (e.g., population effects and different sets of varying effects) remain conditionally independent (Fuglstad et al., 2019, 2020). This leaves a gap in how to jointly regularize multiple components. Priors that behave reasonably for a single set of exchangeable coefficients can produce unintuitive behavior once the model includes multiple components that compete for explained variance.

Grouped predictors refer to situations where the predictor space can be partitioned into subsets, with predictors in each subset sharing some characteristic or relationship (Giraud, 2014). Such grouping arises naturally from multiple sources. For instance, domain knowledge may suggest that variables cluster into meaningful categories. Model structure itself can also induce grouping. In linear models with higher-order interactions, main effects may be considered as one group, two-way interactions another, three-way interactions yet another, and so on. Similarly, generalized additive models can be faithfully represented through interaction linear models where the basis functions for each smooth term constitute a group (Hastie and Tibshirani, 1986; Griffin and Brown, 2017). More broadly, as the number of predictors included in the model increases, groups of related variables arise naturally, making it necessary to consider grouping as an organizing principle (Boss et al., 2023). Exploiting this structure matters for inference since regularization should account for group membership, and we are

interested in how variance is allocated both within and between groups (Buehlmann et al., 2014). Many shrinkage constructions proposed for grouped predictors do not reduce to the standard ungrouped prior as a special case when there is only a single group. This means the grouped formulation is not a proper generalization of the baseline, theoretical properties may be lost, and comparison to standard methods becomes difficult (Xu et al., 2016; Boss et al., 2023). In practice, a recurring feature of grouped shrinkage priors is that the amount of regularization depends on group size. This can be inappropriate when substantive beliefs suggest that heterogeneity of effect sizes varies by group for reasons unrelated to its cardinality. It also misses an opportunity that is often important in applications, which is to learn from the data how strongly shrinkage should be shared within groups and how strongly groups should compete with each other.

Dependencies among predictors, whether induced by multicollinearity in observational data or by design in structured experiments, create a third challenge. Collinearity implies that many coefficient configurations produce similar fitted values, which makes coefficient-level inference sensitive to the geometry of the design (Christensen, 2020; Piironen et al., 2020). At the same time, in structured domains one often has prior information about relationships among predictors that is naturally expressed through dependence. Examples include beliefs that certain coefficients should co-vary, that differences between coefficients should exhibit particular behavior, or that coefficients should have similar magnitudes. The goal is then to construct priors that incorporate covariance-driven structure while still allowing elementwise shrinkage and the ability to represent sparsity. This combination is difficult to achieve with independent coefficient-wise regularization, because independence ignores the geometry induced by correlated predictors and cannot express beliefs about coefficient relationships. Moreover, even when methods exist to incorporate dependence structure, it is often unclear how to select the covariance specification in practice, and requiring practitioners to specify this structure creates additional overhead and potential for misspecification.

To transfer the uncertainty from R^2 to the coefficients, we follow the variance decomposition approach of Zhang et al. (2020). Since we also place a prior on R^2 and perform variance decomposition, our methods can be considered extensions or generalizations of their R2D2 prior. Variance decomposition provides a natural way forward because it separates two conceptually distinct steps. First, placing a prior on R^2 induces a prior on the total variance of the linear predictor. Second, this total variance must be partitioned to determine how each coefficient contributes to it. The first step remains stable across different structural settings, i.e., our beliefs about signal-to-noise ratios do not change when we move from simple regression to multilevel models, grouped predictors, or settings with dependencies. The second step is where structural considerations enter. We modify how the total variance is decomposed to reflect hierarchies, groupings, or dependencies, while keeping the prior on explained variance unchanged. This modularity allows us to develop a unified framework that adapts to the

diverse structural settings discussed above while maintaining a consistent interpretation of R^2 across the different contexts.

In this thesis we develop four complementary priors for structured high-dimensional regression, each extending the variance decomposition framework to accommodate a different form of structure while preserving interpretable, predictive-first specification.

Chapter 2 provides essential background on Bayesian inference and prior predictive distributions, emphasizing the central role of prior specification and motivating the predictive-first approach we pursue throughout this thesis. We review continuous global-local shrinkage priors, characterizing their theoretical foundations including marginal concentration near the origin and tail behavior that enables separation of signal from noise in high dimensions. We show that priors appearing weakly informative at the parameter level can induce highly concentrated and potentially unrealistic distributions on lower-dimensional quantities of interest, motivating the need for direct specification on interpretable predictive summaries. We introduce shrinkage factors and the effective number of nonzero coefficients as tools for understanding how hyperparameters translate into regularization strength and model complexity. Finally, we present the R2D2 prior and variance decompositions, which serve as the foundation for the structured extensions we develop in subsequent chapters.

Chapter 3 introduces the R2D2M2 prior for multilevel models, enabling joint regularization of population-level and varying effects through a variance decomposition that respects hierarchical variance sharing. Coefficients within the same varying-effects term share a variance component, inducing dependence within terms while allowing competition across terms. We use an alternative parametrization of R2-based priors using Gaussian conditionals and establish that this parametrization achieves the same marginal concentration near zero and polynomially decaying tails while being more parsimonious. Parsimony is particularly valuable in multilevel settings where the number of coefficients grows rapidly with grouping structure. We propose a multilevel analogue of effective model size and validate predictive performance and recovery across sparse and dense regimes, including an application in which substantially richer varying-effects structures become feasible.

Chapter 4 demonstrates that the simplex distribution used for variance decomposition is itself a consequential modeling choice. The Dirichlet allocation in R2D2 induces restrictive negative dependence among variance proportions and ties dependence tightly to the mean, limiting the ability to encode substantive beliefs about how terms should compete for explained variance. We introduce generalized decomposition priors (GDR2) based on Logistic-Normal specifications that represent dependence through an explicit covariance matrix while retaining interpretability. We provide practical hyperparameter strategies, including dimension-aware defaults based on concentration of the latent Gaussian norm and a prior-matching procedure that connects Logistic-Normal decompositions to Dirichlet specifications. Simulations show

that the added flexibility can improve predictive and inferential performance, with the largest gains in low signal-to-noise regimes and under strong predictor correlations.

Chapter 5 extends variance decomposition to grouped predictors through the Group-R2 prior. We propose a two-stage allocation that first distributes explained variance across groups and then allocates each group's share across predictors within the group, separating between-group sparsity from within-group heterogeneity in an interpretable way. This structure allows beliefs about group relevance and within-group signal concentration to be specified separately, and it decouples within-group regularization from group size. We characterize the dependence induced within groups, introduce grouped effective model size summaries, and establish strong marginal concentration near zero with polynomial tails while enabling group-level shrinkage. Simulation studies clarify when exploiting grouping improves performance and when simpler specifications suffice.

Chapter 6 introduces dependency-aware shrinkage priors (DASP), which encode structure directly at the coefficient level through correlation matrices while preserving the marginal concentration and tail properties of a chosen base shrinkage prior. This construction separates marginal scaling from dependence and does not impose distributional assumptions on the local scales. We analyze how dependence alters posterior behavior, showing that joint shrinkage can be stronger than under independent formulations. We propose practical approaches to constructing dependence matrices from the design via regularized precision-matrix estimation and extend effective model size to dependency-aware settings. Simulation studies identify when and to what degree incorporating dependence improves prediction and recovery.

Across all four contributions, we characterize marginal and joint shrinkage behavior by establishing concentration near the origin and tail decay rates that govern signal separation in high dimensions (Polson and Scott, 2011; Pas, 2021). To make the connection between hyperparameters and regularization transparent, we study shrinkage factors for all methods and generalize the effective number of nonzero coefficients to the structured settings considered here (Piironen and Vehtari, 2017). We also provide theory-guided recommendations for hyperparameter specification. Our simulation studies and empirical evaluations are designed to identify when each form of structure improves parameter recovery and out-of-sample predictive performance and when simpler specifications suffice, providing practical guidance on when the added modeling complexity is justified. All methods are implemented in Stan (Stan Development Team, 2026) and are available at `https://github.com/jear2412/R2-priors`.

Finally, **Chapter 7** synthesizes the methodological and empirical findings, discusses limitations, and outlines future directions, including extensions to generalized linear models, constrained settings, and causal inference problems where shrinkage must preserve valid effect estimation. We also discuss broader applications of predictive-first prior specification in modeling domains beyond regression.

2 Background

Uncertainty is everywhere and you cannot escape from it. — Dennis Lindley

In this chapter, we establish the conceptual and technical foundations for the methodological contributions developed in the subsequent chapters. We begin with a review of Bayesian inference and its predictive distributions, emphasizing the central role of prior specification. We continue with high-dimensional linear regression, where the challenge of specifying meaningful priors becomes particularly important when the number of parameters grows with or exceeds the sample size. We review continuous global-local shrinkage priors, which provide adaptive regularization through scale-mixtures representations, and we characterize their behavior through marginal distributions, posterior means, origin and tail behaviors, and shrinkage factors. Finally, we introduce the R2D2 prior, which motivates the variance-decomposition approach central to our work by placing a prior directly on the coefficient of determination R^2 and propagating this uncertainty to regression coefficients through a variance allocation mechanism. These elements provide the background for understanding how the methods developed in Chapters 3–6 extend variance decomposition to structured regression settings while maintaining interpretable, predictive-first prior specification.

2.1 Bayesian inference

Bayesian inference treats unknown quantities as random variables and updates uncertainty about them through conditioning (Robert, 2007). Let y denote observed data, and let θ collect the unknown parameters. A Bayesian analysis specifies a prior distribution $p(\theta)$, that encodes beliefs about plausible values of θ before observing the data, and a data model $p(y \mid \theta)$, which describes the distribution of the data conditional on θ. Together, these distributions define a generative process, where we first sample θ from the prior, then draw y from the data distribution given θ (Gelman et al., 2013). After observing y, Bayes' theorem combines both pieces into the posterior distribution (Lindley, 1972; Berger, 1985)

$$p(\theta \mid y) = \frac{p(y \mid \theta)\, p(\theta)}{p(y)} \propto p(y \mid \theta)\, p(\theta), \tag{2.1}$$

which represents updated uncertainty about θ after observing y.

Mathematically, Bayes' theorem is a straightforward consequence of probability theory. However, the interpretation of the posterior as the foundation for inference follows from

Bayesian decision theory (Savage, 1954; DeGroot, 1970; Lindley, 1972; Berger, 1985; Robert, 2007). Under this framework, optimal decisions (such as point estimates, interval estimates, or model comparisons) are those that minimize an expected loss with respect to the posterior distribution. This decision-theoretic foundation provides a principled framework for extracting inferential summaries from $p(\theta \mid y)$.

In modern applications, the posterior $p(\theta \mid y)$ is rarely available in closed form (Gelfand and Smith, 1990; Gelman et al., 2013, 2020b). Practical Bayesian inference therefore proceeds by constructing a computational representation of the posterior that can be interrogated to obtain uncertainty summaries. This representation usually takes one of two forms: (i) a collection of draws $\{\theta^{(s)}\}_{s=1}^{S}$ that approximate sampling from $p(\theta \mid y)$ (as in MCMC), or (ii) an explicit approximating distribution $q(\theta \mid y)$ from a tractable family (as in variational and Laplace-type methods), which may be low-dimensional parametric, highly parameterized, or nonparametric (Robert and Casella, 2004; Blei et al., 2017). Once such a representation is available, posterior quantities of interest can be computed as functionals of it, such as means and other moments, quantiles and credible intervals, tail probabilities, and posterior predictive summaries.

The same logic applies to derived quantities. If $g(\theta)$ denotes a measurable transformation representing a target of substantive interest (e.g., a treatment contrast, a variance component, or a prediction), then probability theory implies that the posterior $p(\theta \mid y)$ induces a posterior distribution for $g(\theta)$. Consequently, posterior uncertainty for $g(\theta)$ can be summarized in the same way as for θ. In practice, this is achieved by applying $g(\cdot)$ to posterior draws $\theta^{(s)}$ (yielding draws $g(\theta^{(s)})$), or by analytically transforming the approximating distribution $q(\theta \mid y)$ when such transformations are available.

2.1.1 Predictive distributions

Beyond the posterior distribution itself, Bayesian inference naturally produces two auxiliary distributions that play distinct but complementary roles in model development, specification, and communication, namely the prior predictive distribution and the posterior predictive distribution.

2.1.1.1 The prior predictive distribution

The prior predictive distribution is defined as

$$p(\tilde{y} \mid \xi) = \int p(\tilde{y} \mid \theta)\, p(\theta \mid \xi)\, d\theta, \tag{2.2}$$

where ξ denotes hyperparameters that govern the prior specification and are specified by the user. We make the dependence on ξ explicit here to emphasize that the prior predictive distri-

bution reflects not only the data model $p(\tilde{y} \mid \theta)$ but also the specific hyperparameter choices made by the analyst. In standard notation, this conditioning on ξ is often suppressed, treating hyperparameters as fixed contextual information (Berger, 1985). Recognizing this dependence is important since varying ξ changes the implied predictive distribution, which allows prior choices to be checked against substantive knowledge about plausible data realizations (Gelman et al., 2020b; Manderson and Goudie, 2026).

A consequence of specifying a prior $p(\theta \mid \xi)$ is that, together with the data model, it implies a distribution over transformations of the data. Let $\mathcal{Y}$ denote the sample space of the observable data, and let $\tilde{y} \in \mathcal{Y}$ denote a replicated outcome (or replicated data) generated from the sampling model, $\tilde{y} \sim p(\tilde{y} \mid \theta)$. For any measurable transformation (or statistic) $T : \mathcal{Y} \to \mathcal{Z}$, where $\mathcal{Z}$ denotes the sample space of the transformed quantity, the prior induces the corresponding prior predictive distribution of $T(\tilde{y})$,

$$p(T(\tilde{y}) \mid \xi) = \int p(T(\tilde{y}) \mid \theta)\; p(\theta \mid \xi)\, d\theta, \tag{2.3}$$

where $p(T(\tilde{y}) \mid \theta)$ is the distribution of the transformed quantity under $\tilde{y} \sim p(\cdot \mid \theta)$. Thus, a prior on θ encodes assumptions not only about parameters but also about the range and structure of data patterns the model considers plausible a priori. Examining this prior predictive distribution reveals whether our prior assumptions are reasonable. Implausible priors can generate datasets that contradict domain knowledge or exhibit unrealistic patterns. This perspective motivates prior predictive checks and, more generally, prior specification strategies that are expressed in terms of interpretable data-level quantities.

In practice, we cannot examine the prior predictive distribution of every transformation $T(\tilde{y})$, as the space of such transformations is vast and many will be irrelevant to the scientific problem at hand. The key is to identify a small set of meaningful summaries (such as the coefficient of determination R^2, the range or variance of responses, group-level means, or other domain-specific statistics) and ensure that $p(T(\tilde{y}) \mid \xi)$ reflects realistic prior beliefs for these quantities. By examining how hyperparameters ξ propagate through the model to influence $p(T(\tilde{y}) \mid \xi)$, we can adjust ξ iteratively until the implied prior predictive distributions align with substantive knowledge. This approach shifts prior specification from abstract parameter spaces to interpretable predictive quantities, making prior specification both more transparent and more defensible.

It is important to maintain perspective on the influence and role of the prior predictive distribution in model building. If the prior predictive distribution does not perfectly match our beliefs about y, provided the discrepancy is not extreme and our beliefs are coherent, this may not be cause for concern. In sufficiently regular parametric models, posterior inferences are often less sensitive to weakly informative prior choices as likelihood information accumulates, though sensitivity can remain substantial for weakly identified parameters and hierarchical

components (Ghosal et al., 2000; Ghosh, 2003; Ghosal and Vaart, 2017). For such models, minor misspecifications of the prior often have negligible impact on posterior conclusions.

However, this reassuring asymptotic behavior breaks down in highly complex models where the number of parameters grows with or exceeds the sample size, as occurs in high-dimensional regression (see Section 2.2) (Ghosal et al., 2000; Song and Liang, 2017; Castillo, 2024). In these settings, the prior retains substantial influence on posterior inference even for large datasets, because each individual parameter receives limited information from the data. The prior predictive distribution can then become a critical diagnostic in which unrealistic prior predictive implications directly translate into unrealistic posterior inferences. For high-dimensional models, careful attention to the prior predictive distribution should not be optional.

The prior predictive distribution also facilitates communication with domain experts and stakeholders who may lack statistical training (Hartmann et al., 2020; Mikkola et al., 2021; Bockting et al., 2024; Manderson and Goudie, 2026). Prior specification is often described in terms of model parameters, even though parameters may not be directly interpretable to domain experts and the mapping from parameters to observable implications can be opaque in complex models. By contrast, experts can usually assess the scale, magnitude, and qualitative features of observable quantities, or of simple summaries of observables (for example, plausible ranges, typical variability, or proportions of variance explained). This makes the prior predictive distribution a practical interface for discussing what the model implies a priori (Gelman et al., 2020a).

Instead of focusing on parameter values, we can present simulated data from $p(\tilde{y} \mid \xi)$ and ask whether they represent plausible outcomes (Gelman et al., 2020b). This moves the discussion to the observable data space, where prior assumptions are easier to inspect and critique. In collaborative settings, prior predictive checks can reveal disagreements about modeling assumptions that might otherwise remain hidden until posterior inference is complete (Gelman et al., 2020b).

While there is substantial literature on prior specification, ranging from objective priors that seek to minimize so-called subjective influence (Berger, 1985, 2006) to informative priors that encode domain expertise (Goldstein, 2006; Gelman et al., 2017), and while this remains an active area of methodological research, we adopt a particular stance in this work. We treat the prior distribution not as an object that can be specified in isolation, but rather as inseparable from predictive quantities it induces (Gelman et al., 2017; Hartmann et al., 2020). From this perspective, prior specification should be guided primarily by understanding prior predictive quantities. In practice, this means examining $p(T(\tilde{y}) \mid \xi)$ for meaningful transformations T and iteratively adjusting ξ until the induced distributions reflect realistic beliefs about observables.

This motivates the reverse procedure; rather than specifying priors on parameters $p(\theta \mid \xi)$ and accepting whatever prior predictive distributions result, we can specify priors directly on interpretable quantities that users can understand and then propagate this uncertainty to the model parameters θ. This inversion of the typical workflow shifts the burden of prior specification from abstract parameter spaces to the observable and prediction space, where domain expertise can reside. Ideally, the resulting prior on θ inherits its structure from our beliefs about predictive quantities, and ensures that regularization at a parameter level aligns with knowledge about the data-generating process (Bhadra et al., 2016; Zhang and Bondell, 2018; Bockting et al., 2024; Manderson and Goudie, 2026).

The principle of examining prior predictive distributions has been advocated in the Bayesian workflow literature (Gelman et al., 2013; Gabry et al., 2019; Hartmann et al., 2020; Gelman et al., 2020b,a). The development of priors that are designed to directly encode beliefs about predictive quantities (rather than merely checking predictive implications after the fact) is still an ongoing area of research. Comparatively fewer approaches make explicit how to construct priors by starting from intuitive beliefs about predictive quantities and then propagating the resulting uncertainty to model parameters. In the context of Gaussian linear models and related additive constructions, notable examples include the work of Fuglstad et al. (2020), who propose hierarchical decompositions of the total variance across the additive terms involved in the random effects and the R2D2 prior of Zhang et al. (2020) which sets a prior over the proportion of explained variance R^2, partitions the total variance and moves the uncertainty to the coefficients.

2.1.1.2 The posterior predictive distribution

After observing data y, the posterior predictive distribution

$$p(\tilde{y} \mid y) = \int p(\tilde{y} \mid \theta)\, p(\theta \mid y)\, d\theta \tag{2.4}$$

describes the distribution of future or unobserved data $\tilde{y}$ conditional on what we have observed. Unlike the prior predictive distribution, the posterior predictive distribution incorporates information from the observed data y through the likelihood (Berger, 1985).

The posterior predictive distribution is used for model validation and prediction (Vehtari and Ojanen, 2012; Vehtari et al., 2016). For model validation, posterior predictive checks compare observed data y to replicated data y^{rep} simulated from $p(\tilde{y} \mid y)$. If the observed data are typical realizations under the posterior predictive distribution, then test statistics computed from y should fall within the distribution of the same statistics computed from y^{rep} (Gabry et al., 2019; Gelman et al., 2020b). Systematic discrepancies between y and y^{rep} indicate aspects of the data that the model does not reproduce, suggesting directions for model revision (Gabry et al., 2019).

For out-of-sample prediction, the posterior predictive distribution quantifies uncertainty about future observations, accounting for both sampling variability and parameter uncertainty. Predictive performance can be evaluated on held-out data using proper scoring rules, such as the log predictive density (Gneiting and Raftery, 2007; Vehtari and Ojanen, 2012; Vehtari et al., 2016). A commonly reported summary is the expected log predictive density (ELPD), defined as the expectation of the log score under the (unknown) data-generating process. In practice, ELPD is approximated using cross-validation. Specifically, the leave-one-out (LOO) estimand for a model $\mathcal{M}$ is

$$\mathrm{ELPD}_{\mathrm{LOO}}(\mathcal{M}) = \sum_{i=1}^{n} \log p(y_i \mid y_{-i}, \mathcal{M}) = \sum_{i=1}^{n} \log \int p(y_i \mid \theta, \mathcal{M})\, p(\theta \mid y_{-i}, \mathcal{M})\, d\theta, \tag{2.5}$$

where y_{-i} denotes the dataset with observation i removed. We estimate $\mathrm{ELPD}_{\mathrm{LOO}}$ using Pareto-smoothed importance sampling (PSIS), which approximates the LOO predictive densities from draws $\theta^{(s)} \sim p(\theta \mid y, \mathcal{M})$ (Vehtari et al., 2016, 2023). PSIS provides a computationally efficient approximation to exact LOO that reuses posterior draws and it offers diagnostics (via Pareto k estimates) that indicate when the importance sampling approximation is reliable. While PSIS-LOO can be used to estimate several predictive criteria, we focus on ELPD because it is based on the log score, a proper scoring rule. Model comparison can then be based on differences in estimated ELPD together with their standard errors, with larger ELPD indicating better expected predictive performance under the log score (Gneiting and Raftery, 2007; Vehtari et al., 2023).

Finally, for applied communication, the posterior predictive distribution provides a direct way to summarize what the model implies for observable outcomes. Posterior predictive intervals and simulated outcome trajectories are often easier to interpret than posterior distributions of latent parameters, particularly when the parameters are not directly meaningful on the outcome scale (Gelman et al., 2020b).

2.2 High-dimensional linear regression

Consider the linear regression model

$$y_i = x_i' b + \varepsilon_i,\ i = 1, ..., n, \tag{2.6}$$

where y_i is the ith response value, x_i is the p dimensional vector of covariates for the ith observation, $b = (b_1, ..., b_p)'$ is the vector of regression coefficients, and $\varepsilon_i \sim \mathrm{Normal}(0, \sigma^2)$ is the residual error. We assume i.i.d errors throughout.

When the number of predictors p is large relative to the sample size n, inference in model (2.6) becomes difficult because the data provide limited information about the individual coefficients.

In particular, if $p > n$ then $X \in \mathbb{R}^{n \times p}$ is not of full column rank, so there exist nonzero vectors v with $Xv = 0$ (Christensen, 2020). Hence b cannot be informed from the likelihood alone, since $X(b + v) = Xb$ for all such v. Even when $p < n$, strong collinearity can make $X'X$ ill conditioned (Hoerl and Kennard, 1970), meaning that the likelihood is informative about some linear combinations of coefficients but only weakly informative about others (Hoerl and Kennard, 1970; Marčenko and Pastur, 1967; Hastie et al., 2009).

From a Bayesian perspective, this implies that the prior plays an especially important role since prior assumptions have a larger impact on the resulting posterior (Ghosh, 2003). All priors can be interpreted as regularizers, for instance, independent Gaussian priors $b_i \sim \text{Normal}(0, \tau^2)$ induce ridge-type penalties, while independent Laplace priors $b_i \sim \text{Laplace}(0, c)$ correspond to LASSO-type penalties (Tibshirani, 1996; Hoerl and Kennard, 1970; Bhadra et al., 2019; Banerjee et al., 2021). The Bayesian framework thus unifies regularization under a probabilistic interpretation, where the prior encodes beliefs about plausible coefficient magnitudes and the posterior balances data fit against prior constraints (Park and Casella, 2008; Bissiri et al., 2016; Murphy, 2022).

High-dimensional regression calls for priors that provide meaningful regularization (Johnstone and Silverman, 2004; Castillo and Vaart, 2012), which can specify what kind of coefficient vectors are plausible by thinking about the typical scale of effects, how to express relationships between p and n, and how to treat structures such as grouping, hierarchies, or dependence among predictors (Boss et al., 2023). Prior design should be centered around the high-dimensional geometry and structure of the problem, ideally allowing practitioners to state their beliefs.

A common structural assumption in high-dimensional problems is the sparsity principle, which posits although the dimension p may be very large, only a handful of the parameters carries substantial signal (Tibshirani, 1996; Buehlmann et al., 2014; Hastie et al., 2015; Tosh et al., 2022). Formally, letting $s = \|b\|_0$ denote the number of nonzero components, the sparsity principle assumes that $s \ll p$ (Pas, 2021). In many applications, sparsity is not merely a modeling decision but a factual claim. For example, in genetics, only a small fraction of genes may be associated with a trait and in signal processing, signals often admit sparse representations (Malone et al., 2011; Bühlmann and Van De Geer, 2011; Wainwright, 2019).

To illustrate the benefits of shrinkage, consider the special case where $X = I$ and $n = p$. This setting is known as the normal means problem (Carvalho et al., 2009)

$$y_i = b_i + \varepsilon_i, \qquad \varepsilon_i \sim \text{Normal}(0, \sigma^2), \qquad i = 1, \ldots, p, \tag{2.7}$$

where we assume for now that the variance σ^2 is known. Despite its simplicity, this model captures the structure of many applications such as estimation of signals corrupted by white noise (Giraud, 2014; Wainwright, 2019), detecting differentially expressed genes (Malone et al.,

2011; Quackenbush, 2002), bankruptcy prediction for publicly traded companies using Altman's Z-score (Altman, 1968), estimating wavelet coefficients in wavelet regression (Wasserman, 2004; Abramovich et al., 2006), and multiple testing procedures in which z scores are approximately normal with unknown means (Efron, 2011) among others.

From the perspective of parameter recovery, the normal means model exhibits a striking phenomenon, known as Stein's phenomenon, that challenged classical intuition. Consider the squared error loss $L(b, \hat{b}) = \|\hat{b} - b\|^2$. The maximum likelihood estimator is $\hat{b}_i = y_i$, which is clearly unbiased. Nevertheless, Stein (1956) showed that $\hat{b}$ is inadmissible for $p \geq 3$, i.e., there exist estimators with strictly smaller risk $\mathbb{E}_b\|\hat{b} - b\|^2$ for all $b \in \mathbb{R}^p$. A canonical example is the James-Stein shrinkage estimator (Stein, 1956, 1981)

$$\hat{b}^{\text{JS}} = \left(1 - \frac{(p-2)\sigma^2}{\|y\|^2}\right) y, \tag{2.8}$$

which dominates the MLE under squared error loss for $p \geq 3$. The improvement is achieved by globally shrinking the vector of observations toward the origin, illustrating how introducing bias can reduce the mean squared error and improve the simultaneous estimation of several parameters. Analogous shrinkage arises from priors that penalize coefficient magnitude (Hoerl and Kennard, 1970; Tibshirani, 1996).

This result came as a surprise to the statistical community because it showed that, even when the coordinates are independent, jointly estimating the vector can reduce overall risk by borrowing strength across components, contradicting the intuition that independent problems should always be treated independently. We clarify that improvement is with respect to the total squared-error risk $\mathbb{E}_b\|\hat{b} - b\|^2$, meaning that it does not imply that every coordinate is estimated better in every sense, and the risk reduction can be unevenly distributed across components and across values of b (Berger, 1985; Lehmann and Casella, 2006). It can be shown that the James-Stein estimator can be expressed as a posterior mode (Takada, 1979; Datta et al., 2025), showing how Bayesian methods implicitly regularise and are able to produce good estimators.

The normal means problem is a natural starting point for developing and evaluating high-dimensional priors. Most priors for high-dimensional regression proposed in the literature are first studied in this setting before being extended to general regression with correlated designs (Castillo et al., 2015; Castillo, 2024).

Returning to regression with a general design matrix X, the normal means model still provides useful intuition; when many coefficients are estimated jointly, shrinkage-type procedures can improve overall accuracy by pooling information across parameters. Bayes' theorem is clear on how to perform inference in these models, but it does not, by itself, determine how the prior $p(b)$ should be specified. Decision-theoretic results imply that, under regularity conditions, a

Bayes rule arising from a proper prior is admissible for the corresponding loss, in the sense that it is not uniformly dominated in risk (Robert, 2007). Admissibility is only a minimal guarantee and does not determine which proper prior yields stable and well-calibrated inference in finite samples, especially in high-dimensional settings where some aspects of b are only weakly informed by the likelihood. Prior specification should thus also be guided by considerations such as alignment with substantive knowledge, interpretability, and computational tractability (Gelman et al., 2020b).

In this thesis, as in related work, we adopt sparsity as a working principle for high-dimensional regression and focus on a family of shrinkage priors known as continuous global-local shrinkage priors (Polson and Scott, 2011, 2012b; Armagan et al., 2011; Bhadra et al., 2016). We consider that this choice is aligned with the considerations outlined above. These priors can encode assumptions about coefficient magnitudes, they are robust in the presence of a few large effects, and they admit computationally effective implementations. They encourage sparsity through strong concentration near zero while still allowing substantial signals through heavy tails. We develop this framework in Section 2.3.

2.3 Shrinkage priors

2.3.1 Global-local formulations

Shrinkage priors constitute a broad class of continuous prior distributions designed to regularize high-dimensional regression coefficients by pulling them toward zero while allowing large signals to escape aggressive shrinkage. Unlike variable selection approaches that assign discrete probabilities to coefficients being exactly zero, such as spike-and-slab priors (Mitchell and Beauchamp, 1988; George and McCulloch, 1993, 1997), continuous shrinkage priors induce soft thresholding in which small coefficients are shrunk heavily toward zero while large coefficients are preserved with relatively little shrinkage (Polson and Scott, 2011). This family of priors avoids a search over the entire space of models and has been extensively studied within the Bayesian community due to strong empirical performance, computational efficiency, and the availability of fast implementations (Pas et al., 2016; Piironen and Vehtari, 2017; Pas et al., 2017b; Makalic and Schmidt, 2016; Bhattacharya et al., 2016).

Global-local shrinkage priors arise from scale-mixtures of normals (Andrews and Mallows, 1974; Barndorff-Nielsen et al., 1982; West, 1987), taking the following form

$$b_i \mid \lambda_i^2, \tau^2, \sigma^2 \sim \text{Normal}\left(0, \sigma^2\lambda_i^2\tau^2\right), \quad \lambda_i \sim \pi(\lambda_i), i = 1, ..., p, \quad \tau \sim \pi(\tau), \sigma \sim \pi(\sigma), \tag{2.9}$$

where λ_i represents local scales unique to each regression coefficient b_i, τ denotes a global scale shared across all coefficients, and $\pi(\cdot)$ represents the corresponding prior distributions.

Scale-mixtures typically belong to one of two classes, namely those that arise from mixing with (generalized) gamma or inverse gamma distributions or mixing with (generalized) beta or inverse beta distributions (Iorio et al., 2025). The former case is characterized by Zhang et al. (2012), whereas the latter has been thoroughly studied by Armagan et al. (2011) and Iorio et al. (2025).

Therefore, different shrinkage priors arise depending on the distributions chosen for λ_i and τ. Popular choices include the Horseshoe (Carvalho et al., 2009), the Normal-Gamma (Brown and Griffin, 2010), Beta mixtures of Gaussians (Armagan et al., 2011), the Dirichlet-Laplace (Bhattacharya et al., 2015), the Regularized Horseshoe (Piironen and Vehtari, 2017), the Beta-Prime (Bai and Ghosh, 2019), and the R2D2 prior (Zhang et al., 2020).

A common reference point is the Horseshoe prior, obtained by taking Half-Cauchy priors on the scales, e.g. $\lambda_i \sim \mathrm{C}^+(0,1)$ and $\tau \sim \mathrm{C}^+(0,\tau_0)$, where $\tau_0 > 0$ represents the expected sparsity of the coefficient vector (Carvalho et al., 2009, 2010; Piironen and Vehtari, 2017). Many later proposals can be viewed as refinements of this construction that target improved behavior in specific regimes while working under the global-local framework (Bhadra et al., 2017, 2019; Tadesse and Vannucci, 2021).

Strong global shrinkage combined with heavy-tailed local shrinkage is a key reason why global-local priors can separate signal from noise. The global scale τ controls the overall strength of shrinkage and, when equipped with a prior, can adapt to the underlying sparsity of the signal (Polson and Scott, 2012b; Pas et al., 2017a), with smaller values inducing stronger shrinkage of noise coefficients. When τ is fixed rather than estimated, it encodes prior beliefs about the sparsity level, often expressed through the expected effective number of non-negligible coefficients (Piironen and Vehtari, 2017; Pas, 2021). In this sense, τ can be interpreted as a sparsity hyperparameter, analogous in spirit to a prior inclusion probability. More generally, sharing a global parameter that quantifies sparsity and can adapt to the unknown signal level is a key ingredient in theoretical guarantees for global-local procedures, including near-minimax risk bounds under sparsity (Carvalho et al., 2010; Pas et al., 2017a). The local scales λ_i allow coefficient-specific departures from this global shrinkage. For coefficients that behave like noise, the posterior concentrates on small λ_i, which reinforces shrinkage toward zero. For coefficients that represent signal, λ_i becomes large, counteracting the global shrinkage so that large effects are only weakly regularized. This combination produces aggressive shrinkage of many small coefficients together with robust behavior for a few large coefficients (Polson and Scott, 2011; Bai and Ghosh, 2021).

In the global-local formulation, recovering signals under sparsity is analogous to robust modeling for outliers (Polson and Scott, 2011). Most coefficients should be strongly shrunk toward zero, while a small subset of exceptional coefficients must be allowed to escape shrinkage. This robustness intuition is closely related to classical robustness arguments, where heavy-

tailed specifications prevent extreme observations from being over-penalized (Dawid et al., 1973; West, 1984; Andrade and O'Hagan, 2011a; O'Hagan and Pericchi, 2012). In global-local priors, the corresponding mechanism is provided by heavy-tailed local scale distributions, which ensure that large signals can be accommodated without being forced back toward zero.

The structure with two scales is not merely convenient but necessary for these priors to separate signal from noise effectively. To see why, consider the simpler alternative of using only a global scale τ, i.e., setting $\lambda_i = 1$ for all i. For example, this subclass of global priors includes scaled Gaussians (ridge regression) with $\tau^2 \sim \text{InvGamma}(\alpha, \beta)$ and the Bayesian LASSO (Park and Casella, 2008). Carvalho et al. (2010) note that in sparse settings the Bayesian LASSO tries to balance risk due to undershrinking noise and to overshrinking large signals, therefore presenting a tension in which estimates of small signals are too big and estimates of large ones are overshrunk. More generally, Bhattacharya et al. (2012) show that global shrinkage priors following fairly general prior families, which include the inverse gamma and Half-Cauchy priors (Gelman, 2006; Polson and Scott, 2012b), can achieve polynomial concentration near the origin to shrink unnecessary noise, but the probability of detecting signals of large magnitude is exponentially small. They conclude that global shrinkage priors are not flexible enough for high-dimensional problems and advocate the use of both global and local scales.

The considerations outlined above suggest two design principles in sparse settings. First, choose $\pi(\lambda_i^2)$ with heavy tails so that large signals can escape shrinkage. Second, choose $\pi(\tau^2)$ with substantial mass near zero so that noise is aggressively shrunk. When both properties are present, global–local shrinkage priors can reproduce key qualitative behavior of the spike-and-slab prior (George and McCulloch, 1993, 1997), while remaining fully continuous and avoiding discrete exploration of the model space (Polson and Scott, 2011; Bhattacharya et al., 2012).

Since the task of λ_i is to discriminate noise from signal, its distribution should prevent overshrinkage of the latter. If the distribution of λ_i has exponential or lighter tails, then true signals are shrunk to zero, regardless of how large the true value is (Polson and Scott, 2011; Bhattacharya et al., 2012; Pas et al., 2016). Priors with these tails face difficulties balancing shrinkage towards zero with leaving large signals unshrunk. Scaled Gaussian priors and the Bayesian LASSO are examples of this. The latter case is particularly notable since the Bayesian LASSO is not able to mimic its frequentist counterpart, showing that sharing structure does not equate to sharing properties across the frequentist and Bayesian frameworks (Bhadra et al., 2019).

The prior on the global scale τ^2 should reflect sparsity levels (Piironen and Vehtari, 2017; Tadesse and Vannucci, 2021). Polson and Scott (2012b) study a wide class of hypergeometric Beta-Prime (Inverted-beta) priors, which contain as a special case the Half-Cauchy prior, which they propose as a default prior. This recommendation rests on two considerations.

First, since the marginal likelihood of the data, as a function of the global scale $m(y \mid \tau^2)$, does not vanish when $\tau^2 = 0$, neither should the prior density $p(\tau^2)$. Otherwise the posterior of τ^2 will be inappropriately biased away from zero. The bias will be most severe near the origin, precisely in the region of parameter space where the benefits of shrinkage become most pronounced. Second, the resulting Bayes estimator under Half-Cauchy priors shows favorable frequentist risk whenever $\|b\|$ is small, showing admissible behavior. Other authors arrive at similar conclusions when studying shrinkage priors, and the distributions of the global scales they choose can be shown to fall within the class of hypergeometric priors (Armagan et al., 2011; Bai and Ghosh, 2021; Zhang et al., 2020). Pas et al. (2016); Ghosh and Chakrabarti (2017); Bhadra et al. (2016) provide guidelines on how to set distributions on the local and global scales to achieve desired behavior, which we synthesize below.

2.3.1.1 Marginal distributions

To compare different shrinkage priors, we study the marginal prior distribution of the coefficients to understand how the choice of local and global scales interact and how their hyperparameters determine the behavior of the prior. Without loss of generality, assume that $\sigma = 1$. The marginal prior distribution of b_i is given by

$$p(b_i) = \int_0^\infty \int_0^\infty \phi(b_i \mid 0, \tau^2\lambda_i^2)\pi(\lambda_i^2)\pi(\tau^2)\, d\lambda_i^2\, d\tau^2, \tag{2.10}$$

where $\phi(\cdot \mid \mu, v)$ denotes the normal density with mean μ and variance v. The flexibility of the scale-mixtures representation allows us to design priors with high concentration around zero and heavy tails by selecting appropriate mixing distributions $\pi(\cdot)$ for λ^2 and τ^2. This integral typically does not have a closed-form solution. We usually make use of approximation theorems to characterize concentration properties near the origin and the tails (Miller, 2006; Olver et al., 2010; Pas et al., 2014; Ghosh and Chakrabarti, 2017).

Rather than choosing $\pi(\lambda^2)$ heuristically, one can restrict attention to mixing distributions with regularly varying tails (Seneta, 1976). This places prior design in a class with tractable asymptotic behavior. Tail indices then provide a simple language for comparing priors, and standard results for regularly varying functions can be applied to derive the induced marginal tails for b (Bhadra et al., 2016).

A function f is regularly varying at infinity with index $\alpha \geq 0$ if f is defined on $[0, \infty)$ and

$$\lim_{x \to \infty} \frac{f(tx)}{f(x)} = t^\alpha, \quad \forall\, t > 0. \tag{2.11}$$

The function f is said to be slowly varying if $\alpha = 0$. Regular variation is preserved under many standard operations, which makes it a convenient class for prior design (Jessen and

Mikosch, 2006). With regard to shrinkage priors, regular variation corresponds to polynomial decay, implying heavy tails which protect large signals from overshrinkage.

Not all common prior families satisfy this property. Gaussian and Laplace priors have exponentially decaying tails and therefore are not regularly varying, which helps explain their tendency to overshrink large signals. The latter case might be surprising, since many tend to think of the Laplace prior as the Bayesian counterpart of the LASSO Bhadra et al. (2019). By contrast, inverse gamma type mixing distributions yield polynomially decaying tails and are regularly varying at infinity (Polson and Scott, 2012a; Pas, 2021; Iorio et al., 2025).

The connection between the tail behavior of $\pi(\lambda_i^2)$ and the marginal prior $p(b_i)$ is established through results on scale-mixtures. Barndorff-Nielsen et al. (1982) show that heavy-tail properties of the mixing distribution carry over to the resulting mixture, and they characterize the resulting tail order of $p(b_i)$ in terms of the tail behavior of $\pi(\lambda_i^2)$. Consequently, by choosing $\pi(\lambda_i^2)$ with slowly varying (or more generally regularly varying) tails, we obtain a marginal prior on b_i whose tails are controlled in a corresponding way.

The design principle that $\pi(\lambda_i^2)$ should have heavy tails can be made concrete through specific distributional forms. Polson and Scott (2011); Ghosh and Chakrabarti (2017); Pas (2021) have shown that setting

$$p(\lambda_i^2) = K \left(\lambda_i^2\right)^{-a-1} L(\lambda_i^2), \quad \lambda_i > 0, \tag{2.12}$$

for some $a > 0$, where K is the normalizing constant, and $L(\cdot)$ is a slowly varying function, yields tail-robust shrinkage behavior (in the sense that strong signals remain unshrunk) and leads to asymptotic Bayes optimality in sparse settings, in the sense of minimax risk and posterior contraction rates in sparse normal means models (Pas et al., 2014, 2016; Ghosh and Chakrabarti, 2017).

Under this choice of $\pi(\lambda_i^2)$, the marginal prior on b_i is the normal variance mixture

$$p(b_i) = K \int_0^\infty \phi(b_i \mid 0, \lambda_i^2)\, (\lambda_i^2)^{-a-1} L(\lambda_i^2)\, d\lambda_i^2, \qquad a > 0. \tag{2.13}$$

By the tail characterization for normal variance mixtures (Barndorff-Nielsen et al., 1982), as $|b_i| \to \infty$,

$$p(b_i) \propto |b_i|^{-2a-1}\, \tilde{L}(|b_i|), \tag{2.14}$$

where $\tilde{L}(\cdot)$ is slowly varying. In particular, $p(b_i)$ has Pareto-type tails with index $2a$.

The parameter a controls tail heaviness directly. Since the Cauchy density satisfies $p(b) = \mathcal{O}(|b|^{-2})$ as $|b| \to \infty$, the marginal prior $p(b_i) = \mathcal{O}\Big(|b_i|^{-2a-1}\tilde{L}(|b_i|)\Big)$ has heavier-than-Cauchy density tails when $a < 1/2$, and therefore offers stronger protection for large signals. The case $a = 1/2$ yields Cauchy-order tails, while $a > 1/2$ produces lighter tails, reducing protection for

large signals and increasing the risk of overshrinkage in sparse settings. The slowly varying factor $\tilde{L}$ can affect the tail magnitude without changing the polynomial exponent $2a + 1$ (Jessen and Mikosch, 2006). Namely, continuous shrinkage priors with polynomial tails are able to achieve posterior contraction rates and when coupled with decision procedures achieve variable selection performance close to those of priors with a point mass (Song and Liang, 2017; Castillo, 2024).

In practice, a is determined by the hyperparameters of the chosen prior distribution on λ_i^2. This allows practitioners to calibrate shrinkage priors based on beliefs about signal strength and sparsity by observing how hyperparameter choices affect a. Iorio et al. (2025) further provide insights on how to compare shrinkage priors based on shrinkage behavior around the origin and their tails.

2.3.1.2 Tweedie's formula and posterior means

In the normal means problem, continuous global-local shrinkage priors admit a useful analytical property. Tweedie's formula (Masreliez, 1975; Polson, 1991; Pericchi and Smith, 1992; Efron, 2011) provides a direct expression for the posterior mean of b_i as

$$\mathbb{E}[b_i \mid y_i] = y_i + \frac{m'(y_i)}{m(y_i)} = y_i + \frac{d}{dy_i} \log m(y_i), \tag{2.15}$$

where $m(y_i)$ is the marginal density of the observation y_i under the prior. The formula can be verified using the identity $(b_i - y_i)\phi(y_i - b_i) = \frac{d}{dy}\phi(y_i - b_i)$. Figure 2.1 illustrates this representation for the Horseshoe prior for several values of the global scale τ.

For global-local priors written as normal scale-mixtures (Andrews and Mallows, 1974), Tweedie's representation holds under mild regularity conditions. Carvalho et al. (2010) show that it suffices that the local scale λ has a proper prior and that $m(y_i)$ is finite for all y_i, even when the induced marginal prior density $p(b_i)$ is unbounded at the origin. This representation underlies much of the theoretical analysis of shrinkage priors, including results on posterior concentration, posterior consistency and near-minimax risk (Carvalho et al., 2010; Polson and Scott, 2011; Datta and Ghosh, 2013; Pas et al., 2016).

Equation (2.15) expresses the posterior mean as the observation adjusted by the score function of the marginal model, $\frac{d}{dy_i} \log m(y_i)$. For global-local priors, this adjustment is negative for $y_i > 0$ and positive for $y_i < 0$, so the posterior mean shrinks y_i toward zero. The magnitude of this adjustment is driven by the shape of $m(\cdot)$, with strong shrinkage for small $|y_i|$ when m is sharply peaked at the origin and much weaker shrinkage for large $|y_i|$ when m has heavy tails. If desired, (2.15) can be combined with an estimate of the marginal density $m(\cdot)$ induced by $p(b)$ to obtain fast empirical Bayes shrinkage rules, since evaluating Equation(2.15) reduces to computing the log-derivative of a one-dimensional marginal.

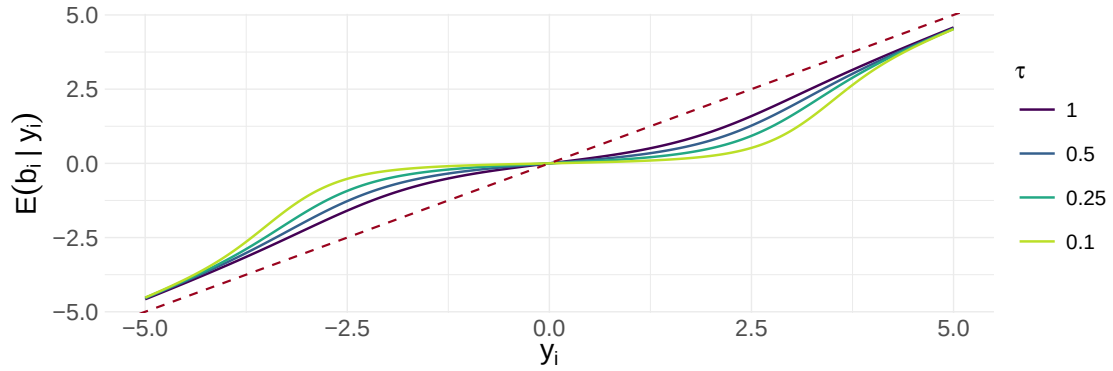


FIGURE 2.1: Posterior mean in the normal means model as a function of y_i under the Horseshoe prior, shown for several values of the global scale τ. The dashed line is the identity map y_i. The deviation from the identity is due to the adjustment term $\frac{d}{dy_i} \log m(y_i)$. Smaller values of τ induce stronger shrinkage toward zero. Shrinkage becomes negligible for large $|y_i|$, so the posterior mean remains close to y_i.

Bayesian robustness may be achieved if $\frac{d}{dy_i} \log m(y_i)$ is bounded as a function of y_i. When this bound converges to 0 as $|y_i|$, we have $\mathbb{E}[b_i \mid y_i] \approx y_i$ ensuring that large signals will not be overshrunk. This mechanism explains the familiar soft-thresholding pattern of global–local shrinkage priors.

2.3.2 Priors on the intercept and variance

Sensible prior choices for the residual variance σ^2 (or equivalently residual standard deviation σ) are provided by Gelman (2006). We consider Half-Student-t priors for σ with ν degrees of freedom and scale η. We propose setting $\eta \approx \text{sd}(y)$, since both the prior expected mean and standard deviation are proportional to η, making this choice weakly informative in the sense that it is scaled to the observed response variability. Alternatively, to enable a Gibbs sampling approach, an inverse gamma distribution $\sigma^2 \sim \text{InvGamma}(c, d)$ with shape and scale parameters c, d respectively could be considered, though care must be taken to avoid overly informative choices (Gelman, 2006).

We set a prior on the intercept $\tilde{b}_0$ implied when all $\mathbb{E}(x_i) = 0$ and then recover the original intercept b_0 after model fitting using a simple linear transformation (Bürkner, 2017; Goodrich et al., 2020). Common priors on $\tilde{b}_0$ are a normal prior with mean $\mathbb{E}(y)$ and a user chosen scale depending on the scale of y as well as Jeffreys prior which is improper flat in this case (Robert, 2007).

2.3.3 Implied priors on lower-dimensional quantities

In many applications, the primary inferential target is not the full high-dimensional parameter vector (b, σ), but a lower-dimensional nonlinear functional of these parameters. Efron (1978)

showed that in such settings, priors that look noninformative for the individual coefficients can induce highly informative priors on nonlinear functionals.

For example, independent priors $b_i \sim \text{Normal}(0, A)$ with $A \to \infty$ may appear flat marginally, yet they can still induce a sharply concentrated prior on the sum of squares $\psi = \sum_{i=1}^{p} b_i^2$. Under this prior one has $\psi/A \sim \chi_p^2$, so $\mathbb{E}(\psi) = Ap$ and $\text{Var}(\psi) = 2A^2p$. In particular, $\psi/(Ap) \to 1$ almost surely as $p \to \infty$ independently of A, so the prior becomes increasingly concentrated around Ap even though each marginal prior on b_i looks flat as $A \to \infty$. Nonasymptotically, an application of the Laurent–Massart concentration inequality shows $\Pr(|\psi/(Ap) - 1| \geq \varepsilon) \leq 2\exp(-c\,\varepsilon^2 p)$ for all $\varepsilon \in (0, 1)$ where c is an absolute constant (Vershynin, 2018). Related phenomena arise for other functionals such as R^2, ratios of coefficients, and maxima. Specifically, R^2 is closely tied to the squared norm of b, and we return to prior specification for R^2 in Section 2.5.

This problem extends beyond improper flat priors. Global shrinkage priors with light tails can strongly regularize individual coefficients, which in turn penalizes functionals that depend on joint deviations of b. In such cases, the induced priors on such functionals can concentrate around values determined by the global scale, and posterior inference can exhibit substantial shrinkage bias for these targets (Bhadra et al., 2017).

The main point is that weak information at the parameter level does not remain weak after a nonlinear transformation. Bhadra et al. (2016) argue that priors with regularly varying tails can mitigate this effect, because heavy-tail behavior propagates more robustly through common nonlinear transformations. This offers an additional motivation for global-local shrinkage priors, since it helps ensure that prior assumptions propagate more coherently to the lower-dimensional quantities we care about.

These considerations have direct implications for prior design. If the goal is to express prior beliefs in terms of predictive quantities or other low-dimensional summaries, it is preferable to use priors whose induced behavior on these functionals remains well calibrated as dimension and scale change.

2.3.4 Hyperparameter specification

Global-local shrinkage priors require choosing hyperparameters that govern both the overall amount of shrinkage and how it varies across coefficients (Tadesse and Vannucci, 2021). For instance, the Horseshoe uses a global scale τ that can be set to reflect an anticipated sparsity level, the Dirichlet-Laplace prior depends on the concentration of the Dirichlet distribution, and the normal-gamma prior is controlled by shape and rate parameters. These choices can strongly influence how aggressively small signals are pulled toward zero and how much large effects are protected.

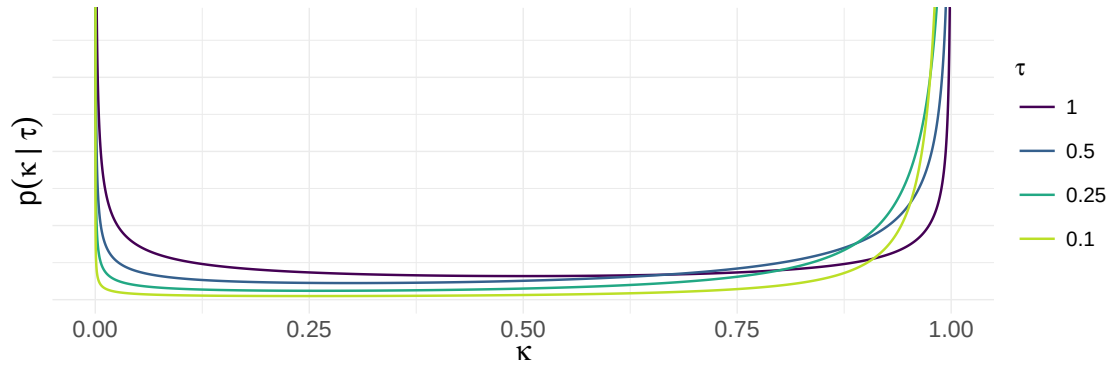


FIGURE 2.2: Density of the Horseshoe shrinkage factor $\kappa \in (0, 1)$ for several fixed values of the global scale τ. The value of τ can be interpreted as the expected proportion of nonzero coefficients. Smaller values of τ concentrate more mass near $\kappa = 1$, corresponding to stronger shrinkage.

The literature proposes default hyperparameter settings with theoretical motivation (Iorio et al., 2025). Such defaults are derived to guarantee specific behavior near the origin and in the tails, and in many cases are linked to optimality criteria such as minimax risk bounds, posterior concentration rates and posterior consistency (Armagan et al., 2013; Pas et al., 2016; Ghosh and Chakrabarti, 2017; Pas et al., 2017a; Song and Liang, 2017). In practice, shrinkage priors are designed to be controlled by only a handful of hyperparameters (without taking into account the hyperparameters involved in the distribution of the intercept and the residual variance). For instance, the Horseshoe prior can be considered as an automatic algorithm, whereas the Dirichlet-Laplace prior can be used by tuning only the concentration parameter of the Dirichlet distribution involved (Bhattacharya et al., 2015; Bhadra et al., 2019). In our simulation studies, we adopt the default hyperparameter values that are proposed in the literature.

Rather than studying the hyperparameters only through their induced marginal distributions on the coefficients, in our work we focus on how they translate into shrinkage behavior. To this end, we introduce shrinkage factors in Section 2.3.5 and the effective number of nonzero coefficients in Section 2.3.6.

2.3.5 Shrinkage factors

To understand how shrinkage priors affect inference, it is useful to examine the posterior mean in the normal means problem. Assume that σ, τ^2 are known. Using Tweedie's formula and Fubini's theorem, the posterior mean can be expressed as

$$\mathbb{E}[b_i \mid y_i] = (1 - \mathbb{E}[\kappa_i \mid y_i])y_i, \tag{2.16}$$

where $\kappa_i \in (0, 1)$ is the shrinkage factor for coefficient i defined as

$$\kappa_i = \frac{1}{1 + \lambda_i^2 \tau^2}. \tag{2.17}$$

Since κ_i lies between zero and one, the posterior mean of each coefficient b_i falls between zero and the observed value y_i. We can think of this as the prior pulling the MLE y_i toward zero, with the strength of that pull determined by κ_i.

The shrinkage factor κ_i quantifies how strongly the prior pushes the estimate toward zero. A value of $\kappa_i \approx 0$ means little shrinkage, with λ_i large and the posterior mean close to the MLE y_i. In contrast, $\kappa_i \approx 1$ means aggressive shrinkage, with λ_i small and the posterior mean close to zero. Carvalho et al. (2010) show that the global parameter τ can be approximated by the average of the κ_i, and they recommend avoiding shrinkage factors that are constrained to lie very close to 0 or 1. Otherwise, τ will not be a faithful measure of the underlying sparsity level, even when it is learned from the data.

By examining the distribution of shrinkage factors under different hyperparameter choices, we can understand the shrinkage behavior of the prior. Ideally, we would like this distribution to place mass where it can accommodate both signal and noise. When p is large and we lack information about which coefficients are active, we need a distribution flexible enough to allow weak shrinkage for some coefficients and strong shrinkage for others.

The Horseshoe prior provides a concrete example. When $\tau = 1$, the shrinkage factors follow a horseshoe-shaped Beta$(1/2, 1/2)$ distribution, which places substantial mass near both $\kappa_i \approx 0$ and $\kappa_i \approx 1$ simultaneously. This bimodal structure allows the prior to adapt to sparse settings, strongly shrinking noise coefficients while leaving signals relatively untouched. This flexibility explains much of the Horseshoe's empirical success. In practice, the global scale will also follow a prior distribution to enable adaptivity (Pas, 2021). We show in Figure 2.2 how the Horseshoe's shrinkage factor κ depends on the global scale τ, with smaller τ pushing the distribution toward $\kappa \approx 1$ and thus enforcing stronger global shrinkage. Armagan et al. (2011) propose a family of shrinkage priors designed through explicit distributions of shrinkage factors, encompassing several well-known shrinkage priors.

2.3.6 Effective number of nonzero coefficients

A joint summary of shrinkage behavior is the effective number of nonzero coefficients (Piironen and Vehtari, 2017), defined as

$$m_{\text{eff}} = \sum_{i=1}^{p} (1 - \kappa_i). \tag{2.18}$$

Since each term $(1 - \kappa_i)$ measures how much coefficient i escapes shrinkage, m_{eff} approximates the number of effectively active predictors in the model. This provides an interpretable summary of model complexity that accounts for partial shrinkage.

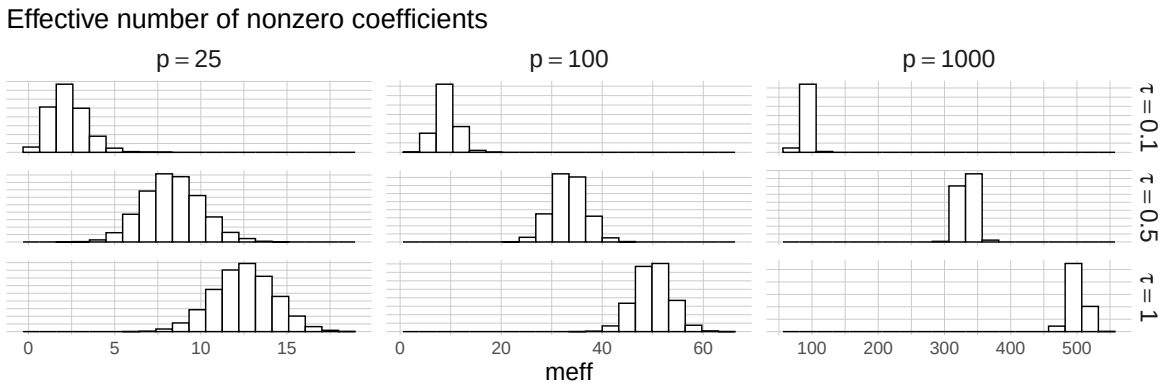


FIGURE 2.3: Prior predictive distribution of the effective number of nonzero coefficients m_{eff} under the Horseshoe prior, shown across number of coefficients p and values of the global scale τ. As τ increases, so does the number of nonzero coefficients m_{eff} in the model, indicating weaker overall shrinkage.

By examining the prior predictive distribution of m_{eff} under different hyperparameter choices, we can understand how hyperparameters translate to expected model size. This facilitates prior elicitation by making the connection between hyperparameter values and model complexity transparent.

2.3.7 Variable selection

Continuous global-local shrinkage priors are well suited for estimation and posterior inference, but they do not produce exact zeros. If a sparse representation is required, an additional variable selection step is needed (Berger, 1985; Robert, 2007; Zhang and Bondell, 2018; Piironen et al., 2020; Tadesse and Vannucci, 2021). This leads to two-step procedures: we first fit a shrinkage model and then apply a decision rule that maps the posterior to a subset of predictors. This separation clarifies the conceptual difference between shrinkage (regularization of estimation) and selection (a discrete decision).

The second step introduces additional choices, such as the selection criterion and any thresholds, which can affect the final selected set. A simple approach is to threshold a posterior summary (e.g., a standardized coefficient estimate) and then include variables exceeding the threshold. Another common strategy is based on shrinkage factors. The decision rule classifies b_i as active if $(1 - \kappa_i) \geq 0.5$ and as inactive otherwise (Carvalho et al., 2010; Datta and Ghosh, 2013). This type of rule has been used for large-scale hypothesis testing, with empirical evidence of good false-positive control in sparse settings (Pas et al., 2017b; Bai and Ghosh, 2021).

A related approach uses marginal credible intervals and selects variables by checking whether zero is excluded. This criterion is aimed at separation rather than interval coverage. For selection, it is enough that intervals for nonzero coefficients exclude zero, even if marginal intervals do not achieve nominal coverage uniformly. For the Horseshoe, Pas (2021) show that

this approach can miss coefficients of intermediate size, while discoveries tend to correspond to true signals.

Bondell and Reich (2012); Zhang and Bondell (2018) propose penalized credible region selection. A full model is fitted under shrinkage and then a nested sequence of posterior credible regions C_α is considered by varying the credible level α. For each region, the sparsest coefficient vector b_α within the region is chosen, producing a sequence of sparse models. This procedure is consistent for model selection in linear models (Zhang and Bondell, 2018), but the required optimization problems can be computationally expensive and may admit multiple solutions, with practical implementations often relying on approximations. Another line of work separates shrinkage and selection by optimizing a prediction-focused loss with an explicit penalty on the size of coefficient vector, rather than thresholding the posterior directly (Hahn and Carvalho, 2015).

Finally, projection predictive variable selection methods provide a prediction-oriented alternative that uses a fitted reference model as a starting point (Goutis and Robert, 1998; Piironen et al., 2020; Catalina et al., 2022; Piironen et al., 2023; McLatchie et al., 2025). The reference model is chosen to give strong predictive performance, and smaller submodels are constructed by projecting the reference model's posterior predictive distribution onto a restricted parameter space corresponding to a subset of predictors. The target is the smallest subset of predictors whose projected posterior predictive distribution remains close to that of the reference model. Model size can be chosen along the projection path using predictive performance estimates such as the ELPD (Vehtari et al., 2023).

2.3.8 Computational methods

Beyond their theoretical properties, shrinkage priors are often convenient computationally. Many can be written as normal scale-mixtures, which enables data augmentation schemes and, in some cases, blocked updates with closed-form conditional distributions (Bhattacharya et al., 2016; Makalic and Schmidt, 2016). These strategies are widely used in MCMC for high-dimensional shrinkage models and exploit the Gaussian structure that appears conditionally in the hierarchy. For an overview of Gibbs and blocked Gibbs samplers in this setting, see Bhattacharya and Johndrow (2021).

In this thesis, posterior sampling is carried out with Hamiltonian Monte Carlo (HMC) and its adaptive variant, the No-U-Turn Sampler (NUTS)(Neal, 2011; Hoffman and Gelman, 2014). These gradient-based methods avoid random-walk behavior and tend to produce draws with low autocorrelation, which can translate into a higher effective sample size per iteration (Brooks et al., 2011; Vehtari et al., 2021). As a result, fewer iterations are often needed to obtain a given level of Monte Carlo accuracy, even in high-dimensional continuous posteriors and hierarchical structures induced by global-local priors. Moreover, this guarantees

consistency across the models we implement and we are not tied to conditionally conjugate parametrizations.

We implement all models in the probabilistic programming language Stan (Stan Development Team, 2026). Stan provides an efficient implementation of NUTS together with automatic differentiation for gradient evaluation. Its sampler adapts tuning parameters automatically during warmup, so in most cases no manual tuning is required. Practical performance still depends on writing the model in a numerically stable way, including appropriate parameterizations such as centered or non-centered forms, and avoiding unnecessary matrix inverses when possible (Zanella and Roberts, 2021). Using a probabilistic programming language also reduces implementation burden and the risk of subtle coding mistakes, while benefiting from community-maintained, well-tested inference software (Gorinova et al., 2019).

2.4 Priors over R^2

Standard global-local shrinkage priors require the specification of hyperparameters that control global and local scales, but the connection between these hyperparameters and predictive quantities such as explained variance is indirect (Zhang and Bondell, 2018; Gelman et al., 2020a). In contrast, R^2 is an interpretable summary of the signal-to-noise ratio measuring the proportion of outcome variance explained by the linear predictor. Variants of R^2 (including pseudo-R^2 measures) are routinely reported and interpreted across disciplines and model classes (Hardin and Hilbe, 2007; Nakagawa and Schielzeth, 2013; Rights and Sterba, 2019), which makes R^2 a convenient target for prior specification (Gelman et al., 2019).

As discussed in Section 2.3.3, priors that appear weakly informative when specified on high-dimensional parameter vectors can induce highly concentrated distributions on lower-dimensional derived quantities. This phenomenon is also exhibited in R^2. Consider independent Normal$(0, 1)$ priors for the regression coefficients b_i and an Exponential(1) prior for the variance σ^2, then the implied distribution of R^2 becomes increasingly concentrated near one as the number of predictors p grows, as shown in Figure 2.4. This is related to the behavior that the sum of squares of the coefficients b_i exhibits as the dimension grows. This behavior reflects unrealistic prior beliefs about the signal-to-noise ratio in most applications and can lead to overfitting. This motivates explicit specification of priors on R^2 rather than relying on implicit distributions induced by coefficient-level priors.

One way to incorporate R^2 into shrinkage prior design is to specify a target distribution for R^2 and then calibrate the hyperparameters of an existing shrinkage prior (e.g., the Horseshoe) so that the induced R^2 distribution matches the target in a divergence sense (Zhang and Bondell, 2018; Bockting et al., 2024). This strategy can be effective, but the mapping from hyperparameters to the implied R^2 distribution is obtained numerically and depends on

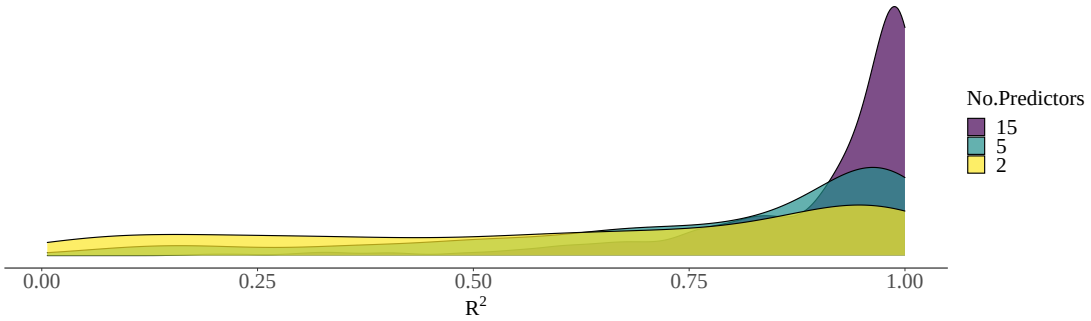


FIGURE 2.4: Implied prior on R^2 for a varying number of predictors in a single-level linear model with Normal$(0, 1)$ priors on the coefficients and an Exponential(1) prior on σ^2. Weakly informative priors imply a highly informative prior on R2.

simulation-based evaluation of the prior predictive behavior. Moreover, because this link is established numerically, the prior's implications for R^2 and for the induced coefficient scales are assessed empirically for the specific calibration setup rather than characterized in general.

We consider a different perspective. Rather than calibrating the hyperparameters of an existing prior, we place a prior directly on R^2 and propagate this uncertainty to the regression coefficients via an explicit transformation. This yields a construction in which R^2 determines a global variance scale for the linear predictor which is then decomposed across coefficients through a simplex allocation. An advantage of this is that it makes the prior suitable for theoretical analysis. The marginal distributions of the coefficients can be derived and their concentration near zero and tail behavior can be studied directly, and theoretical properties such as posterior contraction rates can be established under standard high-dimensional regimes. The resulting prior is the R2D2 prior of Zhang et al. (2020), which serves as the starting point for the prior constructions developed in subsequent chapters. In the remainder of this section we briefly review the R2D2 framework and establish notation for the extensions considered later.

2.5 The R2D2 prior

The R2D2 prior defines a joint prior on the regression coefficients b by first specifying a prior on the coefficient of determination R^2, and then inducing a corresponding prior on b. In this construction, R^2 is defined marginally with respect to a predictor distribution, so the prior on R^2 is specified without conditioning on the realized design matrix. This definition treats x as a random draw to define a design-agnostic prior on explained variance.

Consider a prior for b such that $\mathbb{E}(b) = 0$ and $\mathrm{Cov}(b) = \sigma^2 \,\mathrm{diag}\{\lambda_1^2, \ldots, \lambda_p^2\}$. Assume that the predictors $x \sim H(\cdot)$ with $\mathbb{E}(x) = 0$ and $\mathrm{Cov}(x) = \Sigma_X$ with diagonal elements equal to 1, and that x and ε are independent. Under these assumptions, the variance of the linear

predictor $\mu = x'b$ can be computed using the law of total variance:

$$\mathrm{Var}(\mu) = \mathrm{Var}(x'b) = \mathbb{E}_x\big[\mathrm{Var}_b(x'b \mid x)\big] + \mathrm{Var}_x\big[\mathbb{E}_b(x'b \mid x)\big]\,. \tag{2.19}$$

For the first term we have

$$\begin{aligned}\mathbb{E}_x\big[\mathrm{Var}_b(x'b \mid x)\big] &= \mathbb{E}_x\big[x'\,\mathrm{Cov}(b)\,x\big] \\ &= \mathbb{E}_x\big[\mathrm{tr}\{x'\,\mathrm{Cov}(b)\,x\}\big] \\ &= \mathbb{E}_x\big[\mathrm{tr}\{\mathrm{Cov}(b)\,xx'\}\big] \\ &= \mathrm{tr}\{\mathrm{Cov}(b)\,\mathbb{E}_x(xx')\} = \mathrm{tr}\{\mathrm{Cov}(b)\,\Sigma_X\}\,.\end{aligned} \tag{2.20}$$

Since $\mathrm{Cov}(b)$ is diagonal and $(\Sigma_X)_{ii} = 1$, this reduces to

$$\mathrm{tr}\{\mathrm{Cov}(b)\,\Sigma_X\} = \sigma^2 \sum_{i=1}^{p} \lambda_i^2. \tag{2.21}$$

The second term vanishes since $\mathbb{E}(x'b \mid x) = x'\mathbb{E}(b) = 0$. Thus, $\mathrm{Var}(\mu) = \sigma^2 \sum_{i=1}^p \lambda_i^2$.

We define the marginal coefficient of determination as

$$R^2 = \mathrm{corr}^2(y, x'b) = \frac{\mathrm{Var}(\mu)}{\mathrm{Var}(\mu) + \sigma^2} = \frac{\sigma^2 \sum_{i=1}^p \lambda_i^2}{\sigma^2 \sum_{i=1}^p \lambda_i^2 + \sigma^2} = \frac{\sum_{i=1}^p \lambda_i^2}{\sum_{i=1}^p \lambda_i^2 + 1}, \tag{2.22}$$

which measures the proportion of variance in y explained by the linear predictor μ. Here the equality $R^2 = \mathrm{Var}(\mu)/(\mathrm{Var}(\mu)+\sigma^2)$ follows from $y = \mu+\varepsilon$ with $\mathrm{Var}(\varepsilon) = \sigma^2$ and $\mathrm{Cov}(\mu, \varepsilon) = 0$.

To construct the R2D2 prior we place a Beta prior on the coefficient of determination and parameterize it in terms of its mean μ_{R^2} and precision φ_{R^2}. While Zhang et al. (2020) use the standard shape parameterization, the mean-precision form makes it transparent how prior beliefs about explained variance translate into hyperparameters. The usual shapes α and β are recovered by setting $\alpha = \mu_{R^2}\varphi_{R^2}$ and $\beta = (1 - \mu_{R^2})\varphi_{R^2}$. The prior mean μ_{R^2} is the prior expectation of the explained-variance proportion, and φ_{R^2} controls concentration around this value (Kruschke, 2015).

Next, we define the total prior variance of the linear predictor as $\tau^2 = \sum_{i=1}^p \lambda_i^2$. The relationship between R^2 and τ^2 is given by the deterministic transformation

$$\tau^2 = \frac{R^2}{1 - R^2}, \tag{2.23}$$

which provides a one-to-one mapping between $R^2 \in (0, 1)$ and $\tau^2 \in (0, \infty)$. If $R^2 \sim \mathrm{Beta}(\alpha, \beta)$, then the induced distribution is $\tau^2 \sim \text{Beta-Prime}(\alpha, \beta)$. The Beta-Prime family has regularly varying tails and appears in related constructions for global scales (Polson and Scott, 2012b; Armagan et al., 2011; Bai and Ghosh, 2021). Its tail behavior is governed by the second

shape parameter $\beta = (1 - \mu_{R^2})\varphi_{R^2}$; namely, when $\beta \leq 1/2$ the distribution exhibits heavy polynomial tails, so the prior can support large signals.

The construction of the R2D2 prior proceeds by decomposing τ^2 across the p regression coefficients using simplex-valued proportions $\phi = (\phi_1, \ldots, \phi_p)$ with $\phi_i \geq 0$ and $\sum_{i=1}^{p} \phi_i = 1$. We assign a Dirichlet prior to ϕ with concentration vector α_ϕ (Lin, 2016; Kruijer et al., 2010). The elements of α_ϕ determine how the total variance is allocated across coefficients. We set $\alpha_{\phi,i} = a_\pi$ for all i, which yields exchangeable a priori treatment of the coefficients and reduces the number of hyperparameters.

The concentration parameter $a_\pi > 0$ controls how variance is distributed across coefficients. Small values concentrate mass on a few components (sparse allocation), while large values spread mass more evenly across many components (dense allocation). Each component ϕ_i represents the proportion of total explained variance τ^2 allocated to the i-th coefficient, so that $\lambda_i^2 = \phi_i \tau^2$.

Finally, given the variance components (τ^2, ϕ, σ^2), Zhang et al. (2020) specify the regression coefficients as conditionally independent with double-exponential (Laplace) distributions:

$$b_i \mid \tau^2, \phi_i, \sigma^2 \sim \mathrm{DE}\left(\sigma\sqrt{\frac{\phi_i \tau^2}{2}}\right), \quad i = 1, \ldots, p, \tag{2.24}$$

where $\mathrm{DE}(\delta)$ denotes a double-exponential distribution with mean 0 and variance $2\delta^2$. Since the prior is induced by a prior on R^2 and the total variance is decomposed through a Dirichlet prior, it is referred to as the R^2-induced Dirichlet decomposition (R2D2).

The choice of Laplace conditionals is deliberate, as it yields specific marginal tail behavior and posterior contraction properties, which we discuss below. More generally, any conditional distribution that matches the variance structure

$$\mathrm{Var}(b_i \mid \tau^2, \phi_i, \sigma^2) = \sigma^2 \phi_i \tau^2 \tag{2.25}$$

could be used, although the choice affects the induced marginal shrinkage behavior of b_i. In our work we use Gaussian conditionals for b_i. We discuss the motivation in Chapter 3.

The complete hierarchical specification can be written as:

$$\begin{aligned} b_i \mid \tau^2, \phi_i, \sigma^2 &\sim \mathrm{DE}\left(\sigma\sqrt{\frac{\phi_i \tau^2}{2}}\right), \quad i = 1, \ldots, p, \\ \phi &\sim \mathrm{Dirichlet}(a_\pi, \ldots, a_\pi), \\ \tau^2 &\sim \text{Beta-Prime}(\mu_{R^2}, \varphi_{R^2}). \end{aligned} \tag{2.26}$$

We follow the guidelines of Section 2.3.2 to specify the prior on σ^2 and the intercept, whereas Zhang et al. (2020) suggest inverse gamma priors for σ^2 to enable Gibbs sampling. Zhang et al. (2020) also provide alternative parametrizations of the R2D2 prior, but the formulation we show is the most general and does not impose conditions on the hyperparameters or on how they relate. Importantly, one can show that the R2D2 prior (with either Laplace or Gaussian conditionals) can represent members of the generalized beta mixtures of Gaussians (also known as three-parameter beta distributions) introduced by Armagan et al. (2011).

In this construction, τ^2 serves as a global scale shared across coefficients, while the Dirichlet proportions ϕ act as coefficient-specific local scales. The hyperparameters $(\mu_{R^2}, \varphi_{R^2}, a_\pi)$ control distinct aspects of the induced marginal behavior of b_i. The parameters $(\mu_{R^2}, \varphi_{R^2})$ determine the distribution of τ^2 and therefore the tail behavior of the marginal prior on b_i. Zhang et al. (2020) show that, with $\beta = (1 - \mu_{R^2})\varphi_{R^2}$, the marginal tails satisfy $p(b_i) = \mathcal{O}\Big(|b_i|^{-2\beta-1}\Big)$ as $|b_i| \to \infty$. Above all, smaller values of β correspond to heavier polynomial tails, and the resulting tails can be heavier than Cauchy. They also compare these tails to those induced by the Horseshoe, Horseshoe+, and Dirichlet-Laplace priors, showing that the R2D2 has heavier tails (Carvalho et al., 2010; Bhattacharya et al., 2015; Bhadra et al., 2017).

The Dirichlet concentration parameter a_π governs the behavior near the origin through its effect on how variance is allocated across coefficients. Zhang et al. (2020) show that when $a_\pi < \frac{1}{2}$, the marginal density is unbounded at zero and satisfies $p(b_i) = \mathcal{O}(|b_i|^{2a_\pi - 1})$ as $|b_i| \to 0$. This implies stronger concentration near zero than in priors whose marginal density diverges more slowly, while maintaining heavy tails through the global scale.

The key point is that the R2D2 prior yields polynomial behavior both at the origin and in the tails, at stronger rates than alternative shrinkage priors. In Chapter 3 we discuss that the same types of concentration and tail behavior can be obtained when using Gaussian conditionals in place of Laplace conditionals.

Regarding posterior behavior, Zhang et al. (2020) establish posterior consistency under regularity conditions when p grows with n at rate $p_n = o(n)$, and they derive near-minimax posterior contraction rates. Specifically, the contraction rates in ℓ_2 and ℓ_1 are of order $\mathcal{O}\Big(\sqrt{s\log(p)/n}\Big)$ and $\mathcal{O}\Big(s\sqrt{\log(p)/n}\Big)$, matching the near-minimax rates obtained for spike-and-slab procedures in related settings (Castillo et al., 2015; Song and Liang, 2017; Ročková and George, 2018).

In practice, Zhang et al. (2020) recommend the default choice $a_\pi = \alpha/p$ to facilitate their theoretical analysis and to enable Gibbs sampling implementations, noting that their proofs are tailored to this choice. Other choices might not lead to the same theoretical properties. Under this parameterization, α and β remain to be specified. They suggest $\beta = 0.5$ to obtain Cauchy-like tails, and propose selecting α through a data-size-dependent rule setting $a_\pi = \Big((np)^{\beta/2}\log n\Big)^{-1}$, which implies $\alpha = p\, a_\pi$.

The R2D2 prior enables prior specification in high-dimensional regression through an interpretable predictive quantity, the coefficient of determination R^2. Prior beliefs about the expected explanatory power of the model are encoded through a prior on R^2 and then propagated to the coefficient vector via the induced variance decomposition. This construction, namely the decomposition of τ^2 and the mean-precision parameterization of the Beta prior, serves as the starting point for the extensions developed in this thesis. In Chapter 3 we adapt this framework to multilevel models by allocating variance jointly across overall and varying effects. In Chapter 4 we generalize the Dirichlet allocation by allowing dependence among variance proportions through general decompositions. Chapter 5 develops grouped variance decompositions for settings with known group predictor structure, and Chapter 6 considers dependence specifications at the covariate level.

3 Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior

---

In this work we study prior specification for Bayesian linear multilevel models in high-dimensional settings, where contemporary applications frequently involve many overall and varying effects. Multilevel models rely fundamentally on joint prior distributions to enable partial pooling and stable inference, yet in practice prior specification is often performed independently across different model components (Gelman et al., 2013; Harrell, 2013; Bates et al., 2015). As the number of additive terms grows while the amount of data remains fixed, such independent specifications lead to overparameterized models with poor interpretability, unreliable uncertainty quantification, and degraded out-of-sample predictive performance. In addition, commonly used shrinkage constructions often require the specification of abstract hyperparameters, offering limited joint control over overall and varying effects within a single prior formulation, and do not scale coherently when both types of coefficients are high-dimensional. While global–local shrinkage priors provide an effective regularization strategy for single-level regression models, most existing constructions are tailored to simpler settings and do not directly accommodate the hierarchical structure, variance sharing, and dependence patterns inherent to multilevel models.

The need for joint priors in complex regression settings is well established. They are routinely used for additive components that come with internal structure, such as splines, spatial terms, or monotonic effects (Bürkner and Charpentier, 2020; Wood, 2017; Morris et al., 2019). For multilevel models specifically, joint priors enable partial pooling of varying effects within grouping structures through shared variance parameters (Gelman and Hill, 2006; Fuglstad et al., 2019). However, existing joint prior constructions have been limited in scope: they address regularization within a single random-effects term (e.g., varying intercepts or slopes for one grouping factor) but treat different model components (overall effects, varying effects across multiple grouping factors, and interactions) as mutually independent (Gelman et al., 2013). This component-wise independence becomes increasingly problematic in high-dimensional settings where the cumulative number of parameters grows large relative to sample size, leading to inflated error rates, overconfident uncertainty quantification, and overfitting.

What is missing is a joint prior framework that simultaneously regularizes across all model components while remaining interpretable and computationally tractable. This framework must accomplish three objectives: first, it should provide unified control over the total amount of signal explained by the model rather than requiring separate hyperparameter specifications

for each component; second, it should induce appropriate dependence among coefficients that respects the variance-sharing structure of multilevel models; and third, it should scale appropriately with model complexity.

To address these issues, we introduce the R2D2M2 prior, a global-local shrinkage prior for multilevel models that jointly regularizes the entire multilevel linear predictor through an interpretable prior on the proportion of explained variance R^2. The construction links R^2 to the total variance of the linear predictor μ and allocates this variance across overall and varying-effect components via a Dirichlet decomposition, inducing a joint prior that reflects the variance-sharing structure of random effects. We characterize how the hyperparameters control marginal concentration near the origin, tail behavior, and the implied degree of joint shrinkage. To communicate these effects in practice, we propose interpretable summaries, including shrinkage factors and a multilevel analogue of the effective number of nonzero coefficients. Finally, we validate predictive performance, parameter recovery, and posterior calibration in simulation studies and on benchmark data. We discuss the main ideas in the remainder of the chapter.

To construct our prior distribution, we first define a global coefficient of determination for the multilevel linear predictor μ, which is able to account for the variance contributions of both overall and varying effects

$$R^2 = \text{corr}^2(y, \mu) = \frac{\text{Var}(\mu)}{\text{Var}(\mu) + \sigma^2}, \tag{3.1}$$

and then link it to the total variance $\tau^2 = \text{Var}(\mu)$ via the one-to-one transformation

$$\tau^2 = \frac{R^2}{1 - R^2}. \tag{3.2}$$

As a result, prior information specified on $R^2 \in (0, 1)$ is equivalently transferred to the total variance $\tau^2 \in (0, \infty)$, and vice versa, allowing for direct control of the explained variance. Following the variance decomposition approach detailed in Section 2.5, we decompose τ^2 across overall and group-specific model components using simplex proportions ϕ. The overall regression coefficients receive coefficient-specific variance components, while coefficients associated with the same varying-effect group share variance components, reflecting the repeated-variance structure inherent to multilevel models. Conditional on these variance components, we model all coefficients with Gaussian conditional distributions, matching the standard formulation of random effects in multilevel modeling (Gelman and Hill, 2006). Concretely, for an overall coefficient b_i and a varying coefficient u_{igl} (predictor i, group g at level l) we set,

$$b_i \mid \tau^2, \phi_i, \sigma^2 \sim \text{Normal}(0, \lambda_i^2 \sigma^2), \qquad u_{igl} \mid \tau^2, \phi_{ig}, \sigma^2 \sim \text{Normal}(0, \lambda_{ig}^2 \sigma^2). \tag{3.3}$$

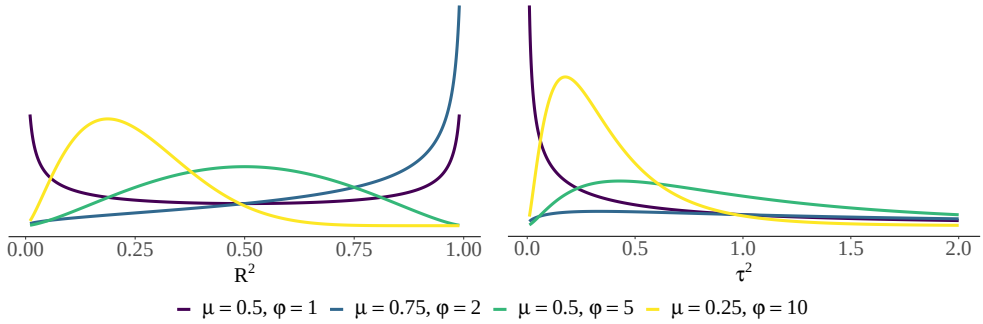


FIGURE 3.1: Exemplary densities of the Beta prior for R^2 (left) and corresponding Beta-Prime prior for τ^2 (right) with varying mean and precision parameters μ, φ respectively.

where $\lambda_i^2 = \phi_i \tau^2$ and $\lambda_{ig}^2 = \phi_{ig} \tau^2$. This variance decomposition yields a joint prior across all overall and varying coefficients, controlled by a small number of interpretable hyperparameters. We place a Beta prior on R^2 parameterized by its mean μ_{R^2} and precision φ_{R^2} (as introduced in Section 2.5), which induces a Beta-Prime prior on τ^2. Depending on $(\mu_{R^2}, \varphi_{R^2})$, we can allocate substantial mass to large values of τ^2, thereby allowing strong signals under global shrinkage. Figure 3.1 illustrates the relationship between the priors of R^2 and the total variance τ^2.

We place a Dirichlet prior on the proportions ϕ and parameterize it through a single concentration parameter $a_\pi > 0$, which governs how strongly variance is concentrated on few terms versus spread across many. Together, $(\mu_{R^2}, \varphi_{R^2}, a_\pi)$ control the overall strength of shrinkage, the tail behavior of the induced global scale, and the degree of sparsity of the implied variance allocation. We show a schematic representation of the R2D2M2 prior in Figure 3.2.

In the R2D2M2 prior, the Dirichlet concentration parameter a_π controls the behavior at zero: for $a_\pi \leq \frac{1}{2}$ the marginal density is unbounded at the origin, inducing aggressive sparsity-promoting shrinkage, whereas for $a_\pi > \frac{1}{2}$ the density is bounded at the origin and the prior transitions smoothly toward denser regimes (Figure 3.3). The R^2 hyperparameters $(\mu_{R^2}, \varphi_{R^2})$ control tail behavior, yielding polynomially decaying tails, including regimes heavier than Cauchy whenever $(1 - \mu_{R^2})\varphi_{R^2} \leq \frac{1}{2}$, that protect large signals from excessive shrinkage under strong global regularization. This implies that estimators derived from our procedures are robust, in the sense that heavy tails protect large signals from excessive shrinkage (Carvalho et al., 2010; Andrade and O'Hagan, 2011b).

Shrinkage priors mainly target sparse regimes and adapt poorly when signals are comparatively dense Bai and Ghosh (2021). Here, sparsity and density are both expressible through the same small set of hyperparameters, so the prior can encode anything from strong sparsity assumptions to fairly dense configurations without changing the model.

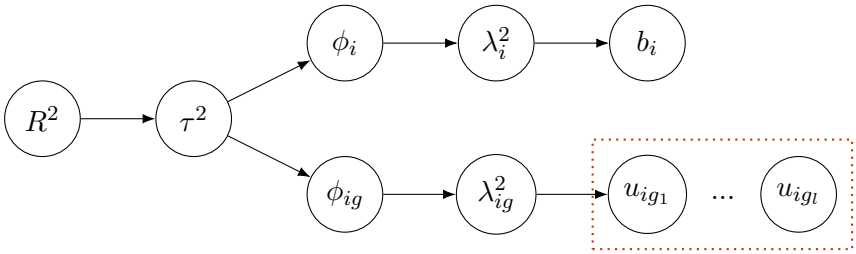


FIGURE 3.2: Schematic representation of the R2D2M2 prior. We set a prior on the proportion of explained variance R^2 and propagate its uncertainty to the regression coefficients through a Dirichlet decomposition of the total variance τ^2. The resulting variance components are assigned to overall and varying effects, inducing a joint prior over all coefficients. Coefficients corresponding to different levels l of the same grouping factor g share a common local variance parameter λ_{ig}^2 for a given predictor i.

A key insight is that we can reproduce the desirable concentration at the origin and the tail-decay rates of the R2D2 prior (Zhang et al., 2020), which uses Laplace conditionals instead of Gaussians. In multilevel modeling, Gaussian conditionals are the default and representing Laplace-based constructions in this form requires a normal scale-mixture with an additional coefficient-specific latent scale, increasing posterior dimension (Andrews and Mallows, 1974; Barndorff-Nielsen et al., 1982; West, 1987). We show that such augmentation is unnecessary to obtain the desired marginal shrinkage properties, which is particularly important in multilevel settings where the number of coefficients can grow rapidly with the number of grouping factors and levels.

We characterize both coefficient-level shrinkage via the distribution of the shrinkage factors κ and joint complexity via a multilevel generalization of the effective number of nonzero coefficients of Piironen and Vehtari (2017) (See Section 2.3.6). We use these two quantities to connect the hyperparameters $(\mu_{R^2}, \varphi_{R^2}, a_\pi)$ to the prior's implied regularization, both for individual effects and jointly across the multilevel structure.

Having established the theoretical properties of the R2D2M2 prior, we now turn to empirical validation. We assess the empirical performance of the R2D2M2 prior through simulation studies that span a range of shrinkage regimes and complexities. We vary $(\mu_{R^2}, \varphi_{R^2}, a_\pi)$ to cover sparse and dense signals, light and heavy tails, and different degrees of variance concentration across model terms, thereby evaluating the prior under realistic and challenging configurations. We focus on evaluating predictive performance and parameter recovery (Benjamini and Hochberg, 1995; Robert, 2007; Vehtari et al., 2023). Across scenarios, we empirically show that the posterior exhibits the expected soft-thresholding pattern: the prior strongly regularizes weak effects while preserving larger signals (Bhadra et al., 2019; Pas, 2021). We further examine posterior calibration using simulation-based calibration (SBC) (Talts et al., 2020) and find well-calibrated inference despite the challenging posterior geometries induced by shrinkage and heavy-tailed signal configurations. Our experiments demonstrate robustness

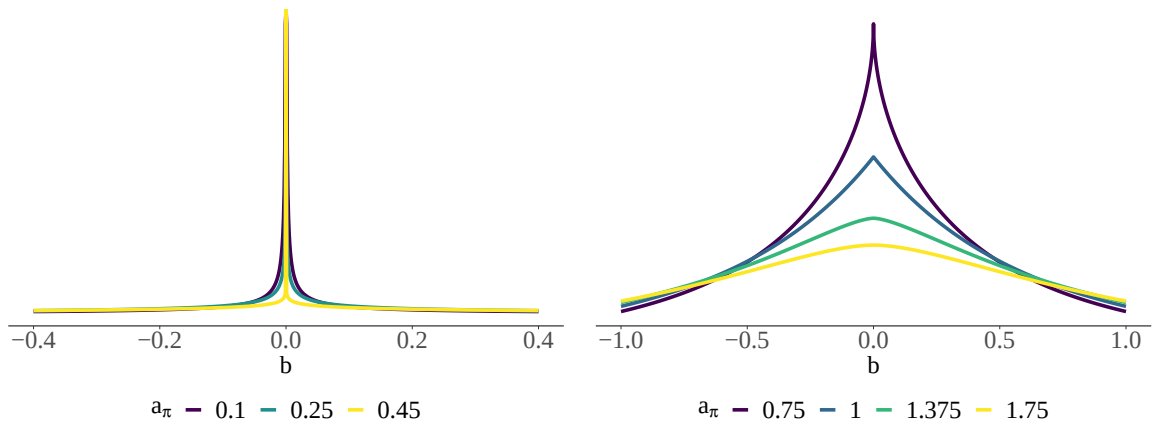


FIGURE 3.3: Marginal prior densities of the regression coefficients for varying values of a_π. The parameter a_π governs the behavior of the density near the origin.

across hyperparameter choices and clarify the trade-offs between sparsity assumptions, tail robustness, and predictive performance.

Finally, we demonstrate the practical value of joint regularization on a benchmark high-dimensional multilevel dataset (Buehlmann et al., 2014; Lin et al., 2020), focusing on how regularization affects signal detection as complexity increases. Whereas previous analyses of this dataset considered only very limited random-effects specifications (not more than 3 varying coefficients), we exemplify how the R2D2M2 prior makes it possible to fit richer varying-effects structures in practice. For practical use, the R2D2M2 prior is available in `brms`, enabling straightforward application (Bürkner, 2017; Stan Development Team, 2026).

In summary, this paper develops a joint shrinkage prior for Bayesian linear multilevel models by formulating regularization at the level of explained variance rather than at the level of individual coefficients. A prior on R^2 provides interpretable control over the global signal-to-noise, while a Dirichlet variance decomposition distributes explained variance across overall and varying effects, yielding a joint prior that shifts uncertainty from predictive quantities to the terms involved in the model. We characterize how the hyperparameters control concentration near zero and tail decay, and show that the same desirable marginal shrinkage properties associated with Laplace-based constructions can be achieved under Gaussian conditionals without inflating posterior dimension. Empirical studies and a real-data application demonstrate stable predictive performance, meaningful recovery behavior, and well-calibrated inference across a range of model sizes, establishing variance decomposition as a principled foundation for prior design in modern high-dimensional multilevel modeling. This variance decomposition perspective provides the foundation for Chapter 4, which generalizes the Dirichlet allocation step by modeling dependence among the variance proportions explicitly through Logistic-Normal decompositions.

4 Generalized decomposition priors on R^2

Most continuous shrinkage priors are formulated within a global–local framework in which regression coefficients are conditionally independent given latent scale parameters. These latent scales are often assumed independent a priori (West, 1987; Armagan et al., 2011; Banerjee et al., 2021). This assumption simplifies computation, but it becomes increasingly restrictive in high-dimensional or structured settings, where one often expects predictors to compete for explanatory power in a structured way. Among existing approaches the Dirichlet–Laplace and R2D2 priors (Bhattacharya et al., 2015; Zhang et al., 2020) introduce dependence among local scales. In these constructions, however, dependence arises implicitly through a Dirichlet decomposition rather than being specified as an object of prior modeling.

This paper builds on the R2D2 and R2D2M2 priors (Zhang et al., 2020; Aguilar and Bürkner, 2023), which decompose the total variance τ^2 of the linear predictor across regression coefficients using a simplex-valued random vector ϕ. This variance decomposition provides an interpretable way to control both the overall strength of regularization and its allocation across predictors. At the same time, the Dirichlet distribution imposes strong structural constraints that limit flexibility in realistic applications. In particular, Dirichlet decompositions enforce strictly negative dependence among the variance proportions and couple the dependence structure tightly to the mean, leaving little room to encode substantive structure in how regression terms share or compete for explained variance (Aitchison, 1986; Barndorff-Nielsen and Jørgensen, 1991). In many applied problems, predictors form groups with shared relevance or hierarchical relationships, and such patterns cannot be represented under the dependence structures induced by a Dirichlet distribution, regardless of how its hyperparameters are chosen (Boss et al., 2023).

We address these limitations by introducing the Generalized Decomposition R^2 (GDR2) prior, whose defining feature is that it models dependence among variance proportions ϕ explicitly. GDR2 treats variance allocation as a central part of prior specification rather than inheriting it implicitly from a default simplex distribution. We achieve this by modeling variance allocation in an unconstrained latent Gaussian space and mapping it deterministically to the simplex using log-ratio transformations (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011). Working in the latent space allows us to specify structured dependence directly through a covariance matrix using standard multivariate tools, and then transfer this dependence to the simplex. We are not aware of earlier continuous global–local shrinkage priors in which dependence structures among variance components are specified explicitly.

Formally, let K denote the number of regression coefficients. We model the variance proportions $\phi \in \mathcal{S}^{K-1}$ by introducing a latent Gaussian vector $\eta \sim \text{Normal}_K(\mu_\eta, \Sigma_\eta)$ and mapping it to the simplex via a differentiable transformation $T : \mathbb{R}^K \to \mathcal{S}^{K-1}$, so that $\phi = T(\eta)$. This induces a Logistic-Normal type distribution on ϕ whose dependence structure is governed by Σ_η. For a symmetric treatment of all components, we use the softmax map

$$\phi_k \;=\; \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}, \quad k = 1, \ldots, K, \tag{4.1}$$

together with an identifiability constraint on η (e.g., $\sum_{k=1}^{K} \eta_k = 0$). Equivalent log-ratio parametrizations commonly used in compositional data analysis are possible, and we provide explicit mappings between these representations (Boogaart and Tolosana-Delgado, 2013; Greenacre et al., 2023).

The remaining construction follows the R^2-based variance decomposition strategy developed in Chapter 3. Crucially, the Logistic-Normal component induces dependence only at the level of variance allocation. Conditional on the variance proportions ϕ and global variance τ^2, the regression coefficients b remain independent Gaussian variables. In this sense, the GDR2 prior can be interpreted as an explicit model for variance competition rather than as a model for correlated regression effects, making it a unique prior in the shrinkage prior framework.

A central methodological contribution of this paper is the set of recommendations for hyperparameter specification for Logistic-Normal decompositions. Although Logistic-Normal models appear widely in topic models and compositional analyses, practitioners often specify priors via empirical Bayes or by placing additional hierarchical priors on latent parameters (Agresti and Hitchcock, 2005; Blei and Lafferty, 2007; Mimno et al., 2008; Chen et al., 2013). Such choices can obscure how hyperparameters translate into behavior on the simplex and can increase posterior complexity. By contrast, our approaches enable direct specification of hyperparameters in terms of interpretable properties of the variance decomposition, allowing users to control prior behavior without introducing additional latent layers or computational overhead.

We first study how (μ_η, Σ_η) determine the induced distribution of ϕ. The mean vector μ_η encodes relative preferences among predictors: setting $\mu_{\eta,k} = 0$ yields exchangeability, whereas deviations allow users to express prior information such as group-level importance by assigning common values within groups. To clarify the role of the covariance structure, we analyze the isotropic case $\Sigma_\eta = \sigma_\eta^2 I$. We interpret the norm $\|\eta\|$ as a global budget that controls how unevenly explained variance can be distributed across predictors. In this setting, $\|\eta\|$ concentrates sharply around $\sqrt{K}\sigma_\eta$, and there exists a constant $C > 0$ such that (Vershynin, 2018)

$$\mathbb{P}\left(\left| \frac{\|\eta\|}{\sqrt{K}\sigma_\eta} - 1 \right| \geq t \right) \;\leq\; 2\exp\left(-\frac{K\sigma_\eta^2 t^2}{2C} \right). \tag{4.2}$$

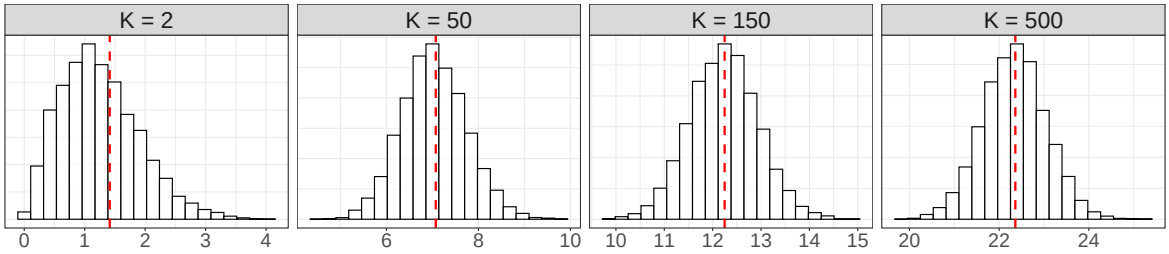


FIGURE 4.1: Implied distribution of the norm of a multivariate normal vector as the dimension K increases. The red line shows the concentration point $\sqrt{K}\sigma_\eta$ for $\sigma_\eta = 1$.

Figure 4.1 illustrates this behavior as K increases. Small values of σ_η shrink η toward the origin and push ϕ toward uniform allocations, whereas larger values permit increasingly uneven allocations and, depending on K, can support sparse variance concentration.

This perspective yields dimension-aware default choices for σ_η. For example, setting $\sigma_\eta = \sqrt{\gamma/K}$ with $\gamma > 0$ keeps $\|\eta\|^2 \approx \gamma$ stable as K grows, while setting $\sigma_\eta = \sqrt{\log(K)/K}$ slows the growth of the norm, preventing excessively large values and reducing variability for large K. Based on our empirical studies, we find the exchangeable choice $\mu_\eta = 0$ together with $\Sigma_\eta = \sigma_\eta^2 I$ and a dimension-aware σ_η to provide a reliable default. We suggest the use of a structured μ_η or non-diagonal Σ_η when substantive dependence information is available.

As a second strategy, we propose a procedure we name prior matching to connect Logistic-Normal decompositions to the more familiar Dirichlet-based specifications. We select Logistic-Normal hyperparameters (μ_η, Σ_η) by minimizing the Kullback–Leibler divergence to a reference Dirichlet distribution $\phi \sim \text{Dirichlet}(\alpha)$. This optimization admits a closed-form solution and introduces no additional computational overhead. Prior matching serves as a practical bridge from Dirichlet-based R2D2 behavior to the more flexible Logistic-Normal framework, while also avoiding numerical instabilities that can arise for Dirichlet priors as $a_\pi \to 0$. Figure 4.2 summarizes how the divergence between a symmetric Dirichlet and its closest Logistic-Normal approximation varies with a_π and K.

We validate GDR2 through simulation studies and benchmark datasets that vary the difficulty of the learning problem (including the signal-to-noise ratio, the correlation structure of the design, and the ratio of sample size to K). Across these settings, the Logistic–Normal decomposition delivers predictive and inferential performance that is on par with, and often superior to, both Dirichlet-based R^2 decompositions and well-established global–local shrinkage priors, with the largest gains in low signal-to-noise regimes and under strong predictor correlations.

In summary, this paper shows that the choice of variance decomposition is a fundamental modeling decision in R^2 based shrinkage priors. By explicitly modeling dependence among

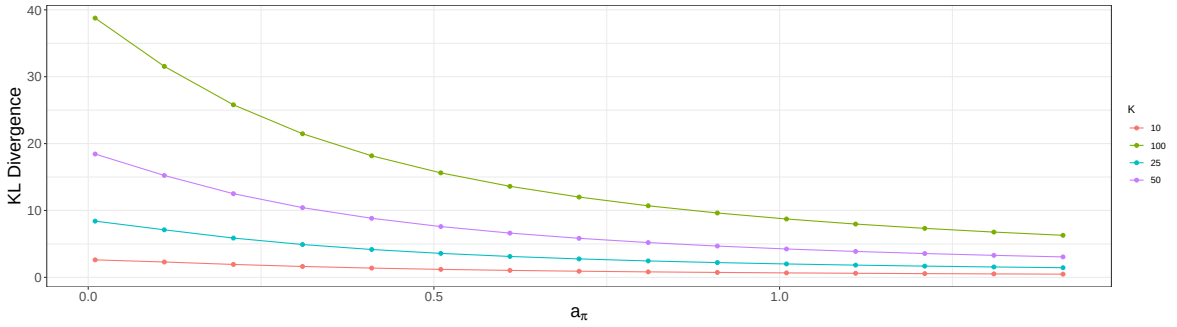


FIGURE 4.2: KL divergence between a symmetric Dirichlet distribution (parameter a_π) and the closest Logistic-Normal distribution as a_π and the number of covariates (K) vary. KL divergence decreases monotonically with a_π.

variance proportions, we achieve predictive and recovery performance that is competitive with reference global–local priors such as the Horseshoe, while retaining the interpretability of variance decomposition based prior specification. More broadly, this work opens a new perspective for shrinkage prior design in which dependence among variance components is specified directly as part of the prior, rather than being fixed implicitly.

Together with Chapter 3, these results extend variance decomposition based approaches to prior specification by enabling structured and interpretable variance competition in high-dimensional regression models. Chapter 5 builds upon these results and shows how we can further enrich R2 priors by allowing them to represent structures at the covariate level.

5 R^2 priors for grouped variance decomposition in high-dimensional regression

---

High-dimensional regression problems often come with natural, externally defined group structures. Examples include genes belonging to biological pathways (Sun et al., 2014), blocks of economic indicators (Mogliani and Simoni, 2021; Kohns and Potjagailo, 2025), or sets of survey items designed to measure a common construct (Bentler and Bonett, 1980; Pauger and Wagner, 2019). Groups can also arise from model structure itself, for instance main effects and their interactions can be treated as separate groups, allowing differential shrinkage between lower-order and higher-order terms (Agarwal et al., 2025). In such settings, the modeling question is not only whether an individual coefficient is relevant, but also whether an entire group is relevant and, if so, whether explanatory power is concentrated in a few predictors or distributed across many within that group (Boss et al., 2023; Tew et al., 2025).

Standard global–local shrinkage priors can be generalized to incorporate grouping by introducing additional group-level scale parameters (Huang et al., 2012; Kang et al., 2016). While several extensions to grouped settings have been proposed, only a few recover the single-group setting as a coherent special case (Xu et al., 2016; Boss et al., 2023). A prominent line of work builds on Bayesian group LASSO–type constructions, but they fail to deliver the marginal behavior often sought in sparse, high-dimensional problems; namely, very strong concentration near zero together with heavy tails. Extensions that do exhibit such properties have no explicit control over within-group shrinkage and depend on group size only (Xu et al., 2016; Boss et al., 2023), which precludes principled prior specification when we expect heterogeneous effect sizes across groups irrespective of dimension. Finally, these constructions provide essentially no guidance on how the imposed group structure translates into predictive, user-facing quantities such as R^2, making it difficult to interpret, calibrate, or compare grouped priors in the predictive space.

Chapters 3 and 4 develop variance decomposition priors that place R^2 at the center of prior design, translating prior beliefs about explained variance into coefficient regularization via a variance decomposition. However, these constructions treat predictors as exchangeable within the decomposition and allocate variance at the level of individual coefficients. Consequently, prior knowledge about the predictive relevance of externally defined groups and about within-group signal patterns cannot be encoded directly.

To address this, we introduce the Group-R2 prior, which extends R^2-based variance decompositions to regression models with externally defined groups of predictors. The key

modification is a two-stage allocation of the total explained variance. We first allocate variance across groups and then allocate each group's share across coefficients within that group. This construction is designed to retain familiar shrinkage features at both the marginal and joint level, while allowing dependence within groups. Additionally, it separates within-group regularization from group size so that within-group heterogeneity is not forced to be a by-product of group size. We connect these behaviors to a small number of interpretable hyperparameters and provide quantities that clarify how the hyperparameters shape shrinkage. Finally, we use simulations to illustrate when exploiting grouping improves performance and when it is neutral or detrimental. The remainder of this chapter develops these results.

We build on the strategy of controlling the total signal variance through $\tau^2 = \frac{R^2}{1-R^2}$ and modify the variance decomposition step to accommodate group structure. Prior specification on R^2 follows the same principles presented in Chapter 3. In Stage I, we introduce simplex weights $\varphi = (\varphi_1, \ldots, \varphi_G)$ to allocate τ^2 across the G groups. In Stage II, for each group g, we introduce within-group simplex weights $\phi_g = (\phi_{g1}, \ldots, \phi_{gp_g})$ to allocate the group-specific share across the p_g coefficients. For parsimony, we consider that the simplex proportions φ, ϕ_g both follow symmetric Dirichlet distributions with concentration hyperparameters a_G, c_g respectively.

Writing b_{gl} for the lth coefficient in group g, we express the induced prior variance allocation through conditional Gaussian distributions,

$$b_{gl} \mid \sigma^2, \tau^2, \varphi, \phi_g \sim \text{Normal}\Big(0,\ \sigma^2\, \varphi_g\, \phi_{gl}\, \tau^2\Big), \qquad g = 1, \ldots, G, \quad l = 1, \ldots, p_g. \tag{5.1}$$

Grouping therefore enters through variance sharing and competition within a global-local Gaussian scale-mixture representation, while retaining the associated computational advantages (Bhattacharya et al., 2016; Makalic and Schmidt, 2016; Tadesse and Vannucci, 2021). The single-level (unstructured) R^2 prior is recovered as a special case by considering a single group.

The two-stage decomposition induces dependence between the prior variance allocations of coefficients within the same group. We show how the strength and sign of this dependence are controlled by the group-level and within-group concentration parameters together with the group size p_g. Importantly, the induced dependence can be either positive or negative, reflecting the balance between a shared group scale and simplex-driven competition. This provides a way to encode prior expectations about within-group signal patterns. Positive dependence corresponds to more distributed signal within groups, whereas negative dependence corresponds to more concentrated, winner-take-most behavior. Figure 5.1 illustrates these effects.

A main contribution of this paper is a theoretical analysis clarifying how the two-stage decomposition separates between-group and within-group shrinkage. Each stage has its own concentration parameter that controls how the variance is allocated: the group-level parameter

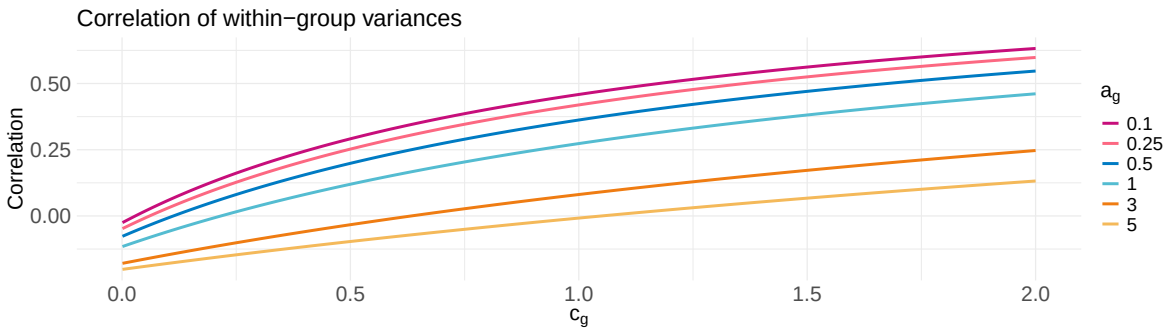


FIGURE 5.1: Correlation between within-group variance allocations $\varphi_g\phi_{gj}$ and $\varphi_g\phi_{gk}$ $(j \neq k)$ as a function of the within-group concentration c_g, for different values of the group-level concentration a_G. Positive correlations correspond to more distributed signal within a group, whereas negative correlations correspond to more concentrated signal. Results are shown for $G = 25$ groups with $p_g = 5$ covariates per group.

a_G governs whether variance concentrates on a few groups or is spread across many, while the within-group parameters c_g govern whether variance concentrates on a few coefficients inside a group or is distributed across the group. Together with the R^2 hyperparameters controlling the induced distribution of τ^2, these parameters determine the marginal behavior at the origin and in the tails. Analysis of the marginal distributions shows that the prior can simultaneously exhibit strong concentration near zero (supporting sparsity) and polynomially decaying tails (protecting large signals). The grouped construction induces both marginal shrinkage of individual coefficients and joint group-wise behavior, implying that the prior can shrink whole groups aggressively. Finally, we also establish that R^2 based priors (grouped or not) can be written in a horseshoe-like form, providing a link to global-local results such as construction of credible regions, which can be transferred to our setting when needed (Pas et al., 2017b; Zhang and Bondell, 2018).

Next, we study how hyperparameters map to interpretable shrinkage summaries. In particular, we introduce a grouped analogue of the effective number of nonzero coefficients (Piironen and Vehtari, 2017) and examine how its prior predictive distribution varies with the group-level and within-group concentration parameters. These summaries provide a communication interface where users can reason about the expected number of active groups and about whether signal within active groups is concentrated in a few coefficients or distributed across many. Figure 5.2 illustrates these effects.

Beyond proposing a structured R^2 prior, we take an explicit stance on investigating when and to what degree including grouping information affects out-of-sample predictive performance (via ELPD-LOO) and parameter recovery (measured by RMSE and False-Discovery rates). In large scale simulations, we vary the group-wise signal pattern, contrasting regimes in which signal within active groups is concentrated in a few predictors with regimes in which it is distributed across many, and we compare against non-grouped R^2 priors with comparable

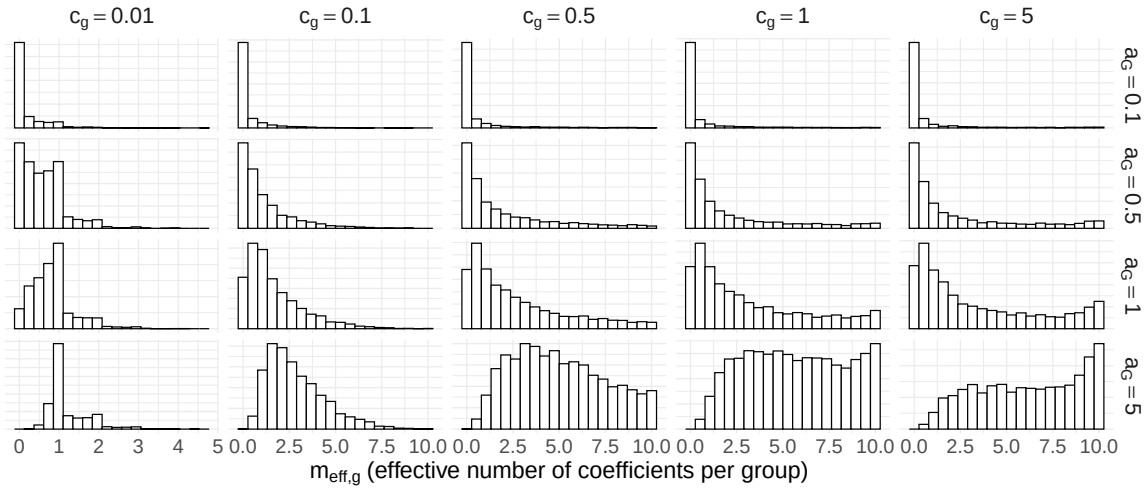


FIGURE 5.2: Prior predictive distributions of the group-wise effective number of nonzero coefficients ($m_{\text{eff},g}$) for different combinations of a_G and c_g. Here, a_G is the concentration parameter of the group-level Dirichlet decomposition (controlling between-group shrinkage) and c_g is the within-group concentration (controlling within-group shrinkage). Rows vary a_G and columns vary c_g. Results are shown for $G = 10$ groups with 10 covariates per group.

overall regularization. Grouping is most beneficial when signals are distributed within active groups and many groups are truly irrelevant, so that group-level shrinkage suppresses noise efficiently. When signals are highly concentrated within groups, the prior can over-shrink the few strong coefficients, and grouping provides little or no benefit.

In summary, the Group-R2 prior extends R^2-based variance-decomposition priors to settings with known group structure by decomposing the prior variance in two stages, across and within groups. This yields structured shrinkage that can improve both prediction and recovery when many groups are irrelevant and signal within active groups is distributed rather than concentrated, with benefits that become more pronounced as dimensionality increases. At the same time, our analysis clarifies the limits of group structure: when signals are extremely concentrated within active groups, group-based shrinkage can overshrink the relevant coefficients. Together with Chapters 3 and 4, this paper expands the variance-decomposition view of shrinkage prior design by showing how prior structure can be aligned with domain structure (groupings) while retaining heavy-tail robustness and interpretable hyperparameter control.

6 Dependency-aware shrinkage priors for high-dimensional regression

---

Dependence is common in high-dimensional regression. Predictors can be strongly correlated, and under collinearity standard procedures often become unstable in how they distribute signal across coefficients. Continuous global-local shrinkage priors regularize high-dimensional coefficient vectors through a global scale and coefficient-specific local scales, and grouped extensions can use externally supplied group structure. However, classical constructions regularize coefficients marginally and do not encode how shrinkage should propagate across correlated covariates. When correlations are strong, this can lead to unstable attribution of signal across collinear predictors and behavior where the posterior concentrates on one representative among highly similar predictors even when signal is shared (Bhadra et al., 2019; Piironen et al., 2020).

Prior work has attempted to incorporate dependence into sparsifying priors. George and McCulloch (1993, 1995) propose dependent spike-and-slab formulations based on inclusion indicators together with dependence linked to active submatrices of the design, but these approaches are costly to fit in high dimensions. More recently, Griffin and Hoff (2024) develop structured shrinkage priors by combining correlated Gaussian variables with random scales. Their framework is developed for coefficient distributions with finite second moments, which excludes some commonly used infinite-variance heavy-tailed shrinkage priors, including the Horseshoe (Polson and Scott, 2011; Bhadra et al., 2017).

Earlier chapters in this thesis focus on variance-decomposition priors, which encode structure through interpretable allocation of explained variance. In this chapter we take a different approach and encode structure directly at the covariate level. We introduce dependency-aware shrinkage priors (DASP), which generalize standard normal scale-mixture constructions by allowing the coefficient vector to have a non-diagonal prior covariance driven by a structure matrix Ω. When the base prior yields strong concentration near zero and heavy-tailed marginals through its choice of τ and λ_i, the proposed generalization can retain these marginal features while changing the joint dependence structure through Ω.

Our construction has three aims. First, it introduces dependence through Ω while keeping the same marginal global-local hyperparameters for τ and λ_i as in the chosen base prior. By constraining Ω to be a correlation matrix, marginal scaling remains governed by the local scales, so heavy-tailed choices for λ_i can be used as usual. Second, it avoids high-dimensional specification of the dependence structure by constructing Ω automatically from the design

matrix via regularized precision-matrix estimation. Third, it evaluates empirically when and to what extent incorporating Ω is useful relative to independent shrinkage.

Let $b \in \mathbb{R}^p$ denote the regression coefficients and let τ and λ_i denote the global and local shrinkage scales as in Section 2.3. We define the dependency-aware conditional prior

$$b \mid \sigma, \tau, \lambda, \Omega \;\sim\; \text{Normal}\Big(0,\, \sigma^2\tau^2 D_\lambda \Omega D_\lambda\Big)\,, \qquad D_\lambda = \text{diag}(\lambda_1, \ldots, \lambda_p), \tag{6.1}$$

together with standard global-local choices for τ and λ_i (e.g., horseshoe-type specifications). The matrix Ω controls dependence and orientation in coefficient space, whereas D_λ preserves the local adaptivity. The usual independent-orientation global-local prior is recovered when $\Omega = I$.

This construction combines two complementary mechanisms for prior specification. Through Ω, it incorporates a priori knowledge of structure by allowing elements of b to covary, encoding similarities and differences among coefficients. The local scales λ_i incorporate a priori knowledge of sparsity via element-wise shrinkage, controlling the relative magnitudes of coefficients. By uniting these mechanisms, DASP spans the range of common structured and sparsity-inducing priors, generalizing standard shrinkage constructions to settings where both marginal regularization and joint dependence are relevant.

We constrain Ω to be a correlation matrix with unit diagonal, which separates marginal scaling from dependence. The local scales λ_i determine coefficient-specific variances, while Ω determines dependence. This avoids confounding between the diagonal of Ω and the local scales and does not require distributional restrictions on λ_i. Under this constraint, the construction preserves the base prior's marginal shrinkage behavior while changing joint shrinkage through the geometry induced by Ω, which affects how regularization acts under collinearity. In particular, we show that dependency-aware constructions place more mass near the origin under the joint prior $p(b)$ than an independent construction $\prod_{i=1}^{p} p(b_i)$ built from the same one-dimensional marginals. We also show that the induced joint prior remains heavy-tailed and places more mass in joint tail regions than the corresponding independent construction.

A practical question is how to specify Ω without turning the prior into a high-dimensional elicitation problem. Rather than placing a prior on Ω (which would introduce many additional parameters) or requiring manual specification (often infeasible), we extract dependence information from the design matrix X. The guiding principle is that X contains information about conditional relationships among predictors through its precision matrix, which can be used to orient shrinkage in directions that respect the dependency structure. This is in the spirit of design-adaptive priors such as Zellner's g-prior, which also uses X to determine the geometry of coefficient regularization (Zellner, 1986).

We construct Ω from the design precision matrix $\Theta_X = \Sigma_X^{-1}$, standardized to obtain a correlation matrix,

$$\Omega \;=\; \mathrm{Cor}(\Theta_X) \;=\; \mathrm{Cor}(\Sigma_X^{-1}).$$

We use the precision matrix rather than the covariance matrix because it encodes conditional dependencies. A zero entry in Θ_X corresponds to conditional independence between the associated predictors given all others. When $p < n$, Σ_X can be estimated empirically. When $p > n$, the empirical covariance is unstable (Marčenko and Pastur, 1967; Ledoit and Wolf, 2020), so we estimate Σ_X using the linear shrinkage estimator of Ledoit and Wolf (2004). This yields a well-conditioned positive-definite estimate with negligible computational overhead and does not require cross-validation or numerical optimization.

The impact of Ω is context-dependent. If the likelihood is strongly informative and the design is well-behaved, the posterior is dominated by the data and dependence in the prior has limited effect. Likewise, if shrinkage is weak, the dependence structure has little room to act. In contrast, Ω matters most when shrinkage is strong and the design is ill-conditioned, because the prior then plays a larger role in stabilizing weakly identified directions. In those regimes, DASP regularizes coefficients jointly along correlated directions rather than independently, which can mitigate instability under collinearity. The effect is amplified when local scales are heterogeneous, because the interaction $D_\lambda \Omega D_\lambda$ induces direction-dependent pooling among coefficients that are shrunk to different degrees.

Geometrically, Ω induces a pooling effect among coefficients associated with correlated predictors. Coefficients are shrunk toward zero as in standard priors, and they are also encouraged to move together when Ω exhibits strong off-diagonal structure. When predictors are highly correlated and share signal, this pooling allows coefficients to reinforce one another rather than competing for posterior mass. Figure 6.1 illustrates the resulting changes in joint prior shape under increasing correlation. To provide an interpretable diagnostic for how dependence affects model complexity, we extend the notion of effective model size m_{eff} to dependency-aware priors. We define

$$m_{\text{eff}}(\Omega) = \mathrm{tr}(I - \kappa_\Omega)\,, \tag{6.2}$$

where κ_Ω is the dependency-aware shrinkage operator. When $\Omega = I$ we recover the usual definition $m_{\text{eff}} = \sum_{i=1}^{p}(1 - \kappa_i)$. As predictor correlations strengthen, effective model size can increase due to pooling among correlated coefficients. The prior distribution of $m_{\text{eff}}(\Omega)$ under the estimated Ω shows how hyperparameter choices and the empirical dependence structure translate into effective complexity. When this diagnostic indicates substantial inflation in model size, stronger shrinkage hyperparameters can counterbalance the effect.

Our simulation studies show that the benefits of incorporating dependence depend on the inferential goal. For predictive performance, improvements are modest to negligible.

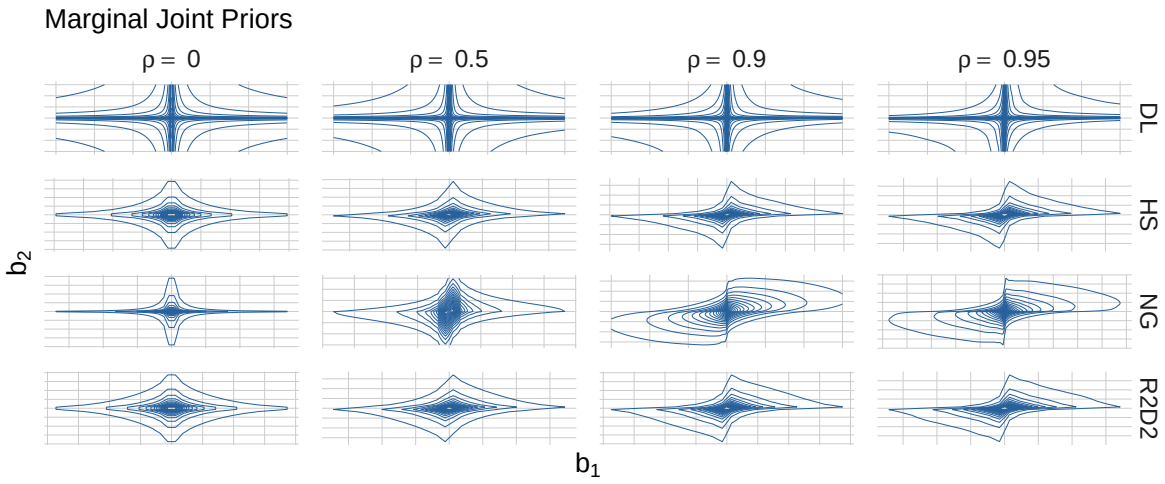


FIGURE 6.1: Monte Carlo approximations of the bivariate joint marginal prior distribution $p(b_1, b_2)$ under different shrinkage priors: Dirichlet-Laplace (DL), Horseshoe (HS), Normal-Gamma (NG), and R2D2, and for various correlation levels ρ. We use default hyperparameters for each prior, with $\sigma = 1$, and $\Omega = (1 - \rho)I + \rho JJ'$, where J is a 2-dimensional vector of ones.

Differences in ELPD tend to be centered near zero, indicating that standard shrinkage priors already achieve strong out-of-sample performance even under substantial collinearity (Bhadra et al., 2019). Parameter recovery is where dependence can help. When groups of correlated predictors jointly carry signal and the constructed Ω reflects the relevant dependence pattern, DASP improves coefficient estimation, with gains most pronounced under high correlation and strong signal-to-noise ratios. When the imposed dependence structure does not align with how signal is distributed across coefficients, incorporating Ω can degrade recovery relative to independent priors. These results show that dependency-aware priors are not a default replacement for independent shrinkage. They are most useful when recovery of shared signal across correlated predictors is a primary goal and when the estimated dependence structure is expected to be informative.

This paper approaches structured shrinkage differently from the variance-decomposition priors developed earlier, which encode structure through allocation of the total variance across terms. Competition for total variance induces dependence in the coefficients. DASP introduces dependence directly at the coefficient level by shaping the joint geometry of b through a non-diagonal orientation matrix Ω. This changes how shrinkage propagates across correlated directions, while retaining the usual global-local control of marginal scales through τ and λ_i. These approaches can lead to similar qualitative behavior in some regimes, but they rely on different mechanisms and offer different points of control. Therefore they provide alternative ways to encode structure, and the choice depends on whether structure is more credibly expressed as variance competition across terms or as pooling along correlated directions in coefficient space.

7 Discussion

This thesis develops a framework for designing joint priors in high-dimensional regression by specifying regularization through predictive quantities rather than through individual coefficients. In settings where the number of predictors is large relative to the sample size, the prior can meaningfully shape posterior inference. The central idea is to reverse the usual direction of prior specification. Instead of placing priors on regression coefficients and then inspecting the implied predictive behavior, we begin with priors on predictive quantities that are easier to reason about and propagate this uncertainty to the coefficients. In particular, beliefs about explained variability are encoded through the coefficient of determination R^2, and a variance decomposition transfers this uncertainty to the regression coefficients. This makes prior specification easier to interpret and expresses regularization on a scale that aligns with common practice. The four papers that comprise this dissertation develop complementary parts of this approach and illustrate how variance-decomposition priors can improve both interpretability and empirical performance.

Variance decomposition provides the unifying technical device, making the allocation of explained variance across model components explicit and controllable. A second theme is that many modern high-dimensional problems come with structure, including hierarchical components, known groupings of predictors, and dependence between effects. Chapters 3–6 extend variance-decomposition-based R^2 priors to such settings while preserving the desirable theoretical properties.

The central insight connecting the four papers is that our predictive-first approach of placing priors on R^2 together with variance decompositions provides a natural route to joint regularization in structured high-dimensional regression. Rather than treating regression coefficients as independent objects to be regularized one-by-one, we target the linear predictor as a whole and make explicit how its explained variance is allocated across model components. This yields priors that are easier to interpret and to communicate with users.

When the parameter dimension is large, we find ourselves in need of a better way to express prior information than coefficient-scale parameters. Not only is the number of scales to specify too high, but the specification might not remain faithful to quantities that we might have interest in. For instance, diffuse independent coefficient priors can imply concentrated and unrealistic beliefs about predictive behavior as dimension grows, even when they appear weakly informative on the coefficient scale (Efron, 1978; Bhadra et al., 2016). Instead, prior specification should be guided by predictive quantities and by aspects of model behavior that are interpretable at the level of the linear predictor and the data. We motivate this design

principle in Chapter 2 and we apply it in the methodological chapters in structured settings, while offering regularization that acts jointly across multiple components.

A first methodological contribution of this dissertation is a set of priors that regularize structured high-dimensional regression models. Chapter 3 develops this idea for multilevel regression through the R2D2M2 prior, which enables joint regularization of overall and varying effects in regimes where both components can be high dimensional. The resulting regularization has two sources, shrinkage induced by the prior and shrinkage induced by multilevel partial pooling. The construction applies to a broad class of multilevel specifications, and we evaluate it across a wide range of simulated conditions, showing stable and desirable parameter recovery and predictive performance even as model complexity increases. The framework we present makes it feasible to fit more complex multilevel models with interpretable joint regularization. We remark that it naturally extends to models that admit a multilevel representation, such as spline-based formulations expressed through basis expansions.

Chapter 4 demonstrates that the variance decomposition used in R^2-based priors is not a neutral choice, since it determines the induced dependence among variance proportions and therefore constrains which forms of joint shrinkage the prior can express. Dirichlet-based decompositions, used so far as a default, impose restrictive dependence patterns among variance shares. We propose generalized decomposition priors that allow richer dependence among variance proportions to be specified directly through a Logistic-Normal construction. This generalization is conceptually important because it elevates variance decomposition from a default construction to an explicit modeling choice. Empirically, the added flexibility translates into improved predictive performance, with the largest gains arising in low signal-to-noise regimes and under strong predictor correlations.

We also explored alternative simplex distributions beyond the Dirichlet, but they proved either too limited in the dependence patterns they could represent, too costly computationally, or yielded no meaningful empirical gains in our experiments. An important contribution of this work is the set of guidelines for specifying hyperparameters in the Logistic-Normal distribution. Prior work using Logistic-Normal models often avoids this step by adding another layer of hierarchy through hyperpriors or by relying on variational inference with restrictive approximation families. By contrast, we provide direct hyperparameter specification strategies that are dimension-aware and interpretable. More broadly, our results suggest that modeling dependence at the level of variance allocation is a viable and principled route for constructing new shrinkage priors.

Chapter 5 extends the same perspective to settings where predictors admit groupings through the Group-R^2 prior. As dimension increases, it is common that covariates are naturally organized into blocks and practitioners can propose such groupings. These groups can reflect external domain knowledge or arise from the model itself (such as treating main

effects and interactions as separate groups). The prior allocates explained variance in two stages, first across groups and then within groups, which allows users to express beliefs about which groups are likely to matter and how signal is distributed within active groups. In this way, group competition is represented at the level of variance allocation rather than through coefficient-wise selection, while still allowing heterogeneity within a group. Although we focus on the setting where grouping information is supplied, the construction is compatible with external group identification when groups are uncertain, since any procedure that proposes a grouping can be used and our prior can then be applied conditional on the resulting groups. Empirically, we find that grouping is most beneficial when signal is distributed across predictors within groups rather than concentrated on a small number of coefficients.

Chapter 6 presents a complementary contribution, in which we treat dependence across coefficients as an explicit modeling choice within shrinkage priors. We develop dependency aware shrinkage priors (DASP) for settings where collinearity affects how signal is represented and where coefficients should be regularized jointly. DASP provides a general construction that can be applied to any global-local shrinkage prior without imposing any distributional assumptions on the local and global scales. DASP generalizes existing priors without changing their intended marginal shrinkage profile, yet introducing dependence induces joint shrinkage that can be stronger than what is obtained under conditionally independent coefficients. Dependence can be specified from substantive knowledge, and when such knowledge is not available we consider data-driven dependence structures extracted from the design matrix. We also characterize when introducing dependence improves recovery and when it can be detrimental. In particular, dependency-aware variants can improve parameter recovery under strong structured collinearity, whereas gains for prediction are often limited when the signal does not align with the correlation structure. This points to a targeted use of dependence modeling, especially when recovery rather than prediction is the primary goal.

The priors that we have designed can be controlled through a small number of hyperparameters independently of the dimension of the problem. This aligns with common practice in the shrinkage-prior literature, which ranges from approaches designed to be essentially automatic to priors governed by only a handful of interpretable hyperparameters (Tadesse and Vannucci, 2021; Iorio et al., 2025). The R2D2M2 prior illustrates this particularly well, since multilevel models can exhibit rapid growth in the number of coefficients and therefore benefit from joint regularization that remains low dimensional at the level of prior specification. When additional knowledge is available, the same hyperparameters can be used to incorporate it in a transparent way. In particular, the hyperparameters governing the variance decomposition provide a direct mechanism for expressing relative importance across components, and therefore for encoding assumptions about where signal is expected to concentrate.

Our priors are deliberately designed to retain the shrinkage behavior required in high-dimensional regression, while expressing assumptions on an interpretable predictive scale.

Achieving both goals simultaneously requires careful construction, since placing priors on R^2 alone does not automatically guarantee desirable shrinkage properties. One might expect that reversing prior specification would introduce inefficiencies or require careful tuning to match the performance of methods designed explicitly for coefficient-wise shrinkage. Our design process motivated specific distributional choices and led to default hyperparameter choices that are grounded in theory and validated empirically. These defaults perform on par with well-established methods while making the predictive implications of prior choices transparent.

The defaults we propose reflect generic assumptions about sparsity, signal strength, and the regularization behavior typically required in high-dimensional settings. These defaults support practical use by providing a stable starting point. When domain knowledge is available, practitioners should tailor hyperparameters to encode problem-specific beliefs rather than relying solely on defaults. The advantage of this approach is not that defaults eliminate the need for thoughtful prior specification, but that they allow practitioners to focus their modeling effort on choices, such as which predictors to include, what grouping structures are meaningful, or how strong regularization should be, rather than navigating a prohibitively large number of abstract hyperparameters whose implications might be difficult to assess.

We introduce summaries that make the implied regularization behavior explicit and interpretable. Coefficient-level shrinkage is characterized through shrinkage factors, while model-level shrinkage is quantified via the effective number of nonzero coefficients. Together, these summaries translate abstract hyperparameters into concrete statements about sparsity and model complexity, enabling practitioners to assess whether a proposed prior implies realistic regularization patterns and plausible model size before observing any data. We use these diagnostic tools consistently across all priors studied to provide a common framework for calibrating hyperparameters and comparing the implications of different structural assumptions. Our goal is to make prior specification transparent enough that users can verify how their choices encode realistic beliefs and adjust them when domain knowledge suggests otherwise.

When practitioners approach high-dimensional regression problems, the choice among the structured priors developed in this dissertation should be guided by the structure they believe is present in the data and by the inferential goal. If the model is naturally multilevel, then the R2D2M2 prior provides a principled way for joint regularization. If predictors admit meaningful groupings and the goal is to learn how signal is distributed across groups as well as within groups, then the Group-R^2 prior offers a direct way to encode and study that structure, and it can improve both prediction and recovery when the grouping reflects how signal is organized. If collinearity is strong and coefficients are expected to share information in a way that aligns with the predictor correlation structure, then DASP is a natural option, particularly when parameter recovery is the priority. Beyond these structural considerations, practitioners should calibrate the overall degree of sparsity and regularization using prior

predictive summaries, and the prior predictive distribution of effective model size provides a practical tool for checking whether a proposed prior implies a plausible level of complexity before fitting the model to data.

Importantly, these approaches are not mutually exclusive. The variance decomposition framework allows combinations that exploit multiple forms of structure simultaneously. For example, Group-R2 can be applied within the overall effects component of a multilevel model. Similarly, generalized decompositions can be used with the other priors that we have presented. The modular nature of variance decomposition means that structural assumptions can be layered when the problem warrants it.

We do not treat structured shrinkage as a uniformly superior replacement for standard baselines. We have designed the empirical studies in this dissertation to characterize when, and by how much, incorporating structure is beneficial, and when it is not. This motivates a deliberate use of structured priors. The choice to use them should also reflect the inferential goal, since some constructions primarily improve parameter recovery while offering limited gains for prediction, and vice versa. Structured models can also present an additional computational cost, so their use should be justified by gains that matter for the application. In practice, a simpler baseline that is fast and reliable can be preferable to a substantially more expensive model whose improvements are small or unstable. Our guidelines should help practitioners to choose an appropriate level of modeling complexity and to determine which direction to pursue.

The variance decomposition framework we have developed addresses several challenges in high-dimensional prior specification, but important limitations remain that both define the scope of this work and suggest productive directions for extension.

A current limitation of our approach is that the variance decomposition framework is restricted to linear models. R^2 is naturally defined as a proportion of explained variance in the Gaussian case, but this notion becomes ambiguous in generalized linear models. Although the community has proposed several R^2 measures for GLMs, there is no canonical variance decomposition that yields a single, universally accepted notion of explained variance (Hardin and Hilbe, 2007; Nakagawa and Schielzeth, 2013; Rights and Sterba, 2019). Different definitions target different concepts and operate on different scales, so the same model can yield materially different R^2 values depending on the chosen metric. Some proposals are not constrained to lie in the unit interval, which weakens their interpretation as a proportion of explained variation. Others are defined via likelihood or deviance comparisons and quantify relative fit rather than a variance share, with their meaning depending on the assumed likelihood and baseline model.

This shows that a principled R^2 definition with sufficient generality is needed for GLMs. We are currently developing a pseudovariance approach that approximates the GLM likelihood with a Gaussian distribution whose mean and variance can be expressed analytically, allowing

construction of an R^2-like quantity whose scale depends on the link function and family. The advantage of this approach is that R^2 remains defined in terms of the variance of the linear predictor, which results from approximating the likelihood at the level of the response variable (Gelman et al., 2013). Preliminary results are promising and we have been able to establish theoretical properties of the prior. We note that in comparison to linear models, research on shrinkage priors specifically for GLMs is less extensive and requires further development (Jiang, 2007; Tang et al., 2017; Jeong and Ghosal, 2020; Bai, 2025).

Extensions to Gaussian processes and neural networks present deeper challenges for variance decomposition approaches. For Gaussian processes, variance-based measures may be less natural than functionals that capture process behavior, such as boundary crossing events, excursion set sizes, or smoothness measured through derivatives (Rasmussen and Williams, 2006). For neural networks, extreme nonlinearity and high-dimensional parameter spaces may require fundamental departures from variance decomposition and we might instead propose general measures of complexity (Ghosh et al., 2019; Izmailov et al., 2021).

Each of the individual methods we have developed also has specific limitations. The R2D2M2 prior assumes diagonal covariance structures for random effects. While we have already developed the necessary theory for non-diagonal covariances, interpretation and posterior sampling remain challenging due to the need to preserve variance constraints while allowing correlation. The current diagonal version should recover coefficients adequately in most applications, but full correlation structures would allow more flexible modeling of dependencies among varying effects within groups.

For the generalized decomposition priors in Chapter 4, we have specified the hyperparameters of the Logistic-Normal distribution either manually or through prior matching to a reference Dirichlet distribution. These strategies provide practical guidance, but procedures that extract information from the design matrix or pilot data to inform the mean vector and covariance matrix of the Logistic-Normal could improve specification when substantive prior knowledge is limited. More general dependency structures on the simplex, such as those based on copulas, remain unexplored but could offer additional flexibility beyond the Logistic-Normal family (Nelsen, 2006).

The same variance-allocation viewpoint also supports priors that encode structure at the group level. The grouped variance decomposition in Chapter 5 assumes that predictor groups are known and supplied by the practitioner. When group structure is uncertain or unknown, the prior could in principle be coupled with procedures that discover groupings. However, this introduces additional complexity and uncertainty propagation challenges that we do not address (Banerjee et al., 2021). In practice, we recommend that groups be grounded in domain knowledge when possible.

For DASP in Chapter 6, we chose to estimate the dependence structure from the design matrix using regularized covariance estimators. More sophisticated Bayesian methods for learning dependence structure (such as graphical Horseshoe or structured covariance learning (Li et al., 2019; Banerjee et al., 2021; Lee et al., 2022)) could improve performance when the design matrix provides limited information about the true coefficient dependencies, but would substantially increase computational burden. DASP is generally more expensive than baseline independent shrinkage priors due to the complex posterior geometries induced by structured covariance matrices even with the dependence structure fixed. Full Bayesian inference over the dependence structure would compound these computational challenges, and our empirical results suggest that even with known dependence structure, DASP provides limited predictive gains in most scenarios. Given these computational costs and limited empirical gains, the most productive direction for improving DASP is not more sophisticated dependence learning, but rather computational innovations that make the existing approach more tractable, such as low-rank approximations of the structured covariance matrix or alternative posterior parameterizations that improve sampling geometry.

Beyond these method-specific limitations, a broader challenge is that the methods we have proposed do not yet accommodate constraints on how coefficients are related, such as positivity constraints, sum-to-zero constraints, or more general convex constraints. While constrained estimation for Gaussian distributions is well-studied, directly constraining the coefficients in R^2-based priors creates the fundamental problem that constraints on the coefficients would alter the variance of the linear predictor in ways that break the R^2 definition and the variance decomposition structure. To preserve the R^2 framework, we propose shifting constraints to the latent space in the generalized decomposition prior. The coefficients inherit their structure from the latent variables through the variance decomposition, so constraints on the latter propagate to the coefficients while preserving the decomposition and keeping R^2 well defined. The main difficulty is computational. Sampling from high-dimensional truncated multivariate normal distributions can be challenging, especially when the truncation region is complex. We are currently addressing this using proximal MCMC methods that relax the constraints (Zhou et al., 2024). One immediate application is interaction modeling, where it is often desirable to enforce hierarchical constraints in which an interaction is only active when its constituent main effects are active (Griffin and Brown, 2017). Preliminary results suggest that variance decomposition priors can be extended to accommodate such structured constraints in a practical way and are better options than standard baselines.

The empirical evaluations in this thesis rely primarily on simulated data and a limited number of real-data examples. While simulation is essential for controlled comparison and for isolating the effects of specific prior choices, broader empirical validation across diverse application domains would strengthen the evidence for practical utility. Collaborations with domain scientists in fields such as genomics, economics, and social sciences could reveal

additional challenges and opportunities for refinement. Encouragingly, the methods we have developed have been adopted by the Stan community, with R2D2M2 and related priors implemented in the widely-used `brms` package (Bürkner, 2017). This suggests that the predictive-first specification philosophy resonates with practitioners, though systematic evaluation of how these priors perform in diverse real-world applications remains an important direction for future work.

An important extension of the predictive-first philosophy is to causal inference, where the inferential target is not prediction but estimation of causal effects. In high dimensions, priors that are designed primarily for prediction can induce a strongly informative implicit prior on a causal estimand. They can also favor regimes with limited confounding by shrinking covariates that contribute more to treatment assignment than to outcome prediction. This can introduce unintended bias in causal effect estimates even when predictive performance is strong (Linero, 2017, 2023, 2024). A natural direction is to specify priors directly on causal quantities of interest and propagate that uncertainty to model parameters, in the same spirit as R^2-based priors. By placing prior structure on the causal estimand itself, or on components that determine it, we could align prior information with the inferential goal. Developing such causal-functional priors could make Bayesian causal analyses more transparent and more robust in high-dimensional settings. This direction connects to broader questions about when shrinkage priors preserve valid causal inference and how to design procedures that retain regularization benefits while recovering reliable causal estimates (Hahn and Carvalho, 2015; Caron et al., 2022).

A broader and more ambitious direction is to move beyond R^2 and use predictive-first prior specification for other interpretable predictive quantities. Existing work on prior predictive calibration and simulation-based elicitation provides a general framework for constructing priors by targeting user-chosen summaries of the prior predictive distribution (Bockting et al., 2024; Manderson and Goudie, 2026). The main idea is to choose a target predictive summary, such as a predictive interval width, a set of predictive quantiles, a calibration curve, or a task-specific accuracy measure defined under a fixed decision rule, and then find hyperparameters so that the induced prior predictive distribution reflects practitioner beliefs about that summary.

This reverse engineering of priors from predictive targets is conceptually straightforward but technically demanding, especially in high dimensions where the mapping from hyperparameters to predictive behavior can be highly nonlinear. A key challenge is to connect the optimization procedure to the expected theoretical behavior of the prior, so that the resulting hyperparameter choices preserve the intended regularization properties. This suggests that calibration should be paired with theory-aware restrictions or criteria so that predictive matching does not come at the expense of the shrinkage behavior we rely on. This requires careful choices of discrepancy measures that define a useful optimization landscape and algorithms that can navigate it reliably (Rényi, 1961; Gretton et al., 2012; Murphy, 2022). A further benefit is that

this approach would allow us to incorporate multiple predictive summaries simultaneously, with the prior resolving potentially competing requirements into a coherent joint distribution.

The same principle extends naturally beyond regression. Survival models provide natural examples, since we can work with quantities such as median survival time, survival probabilities at clinically relevant times, or cure fractions, which are more interpretable than hazard parameters (Klein and Moeschberger, 2003). Spatiotemporal models provide another natural setting, where priors are commonly placed on covariance parameters or precision matrices that are difficult to interpret directly, while practitioners often have clear beliefs about correlation ranges, smoothness, or the expected magnitude of variation across space (Cressie and Wikle, 2011). We could also extend the approach to latent variable models, where priors on loadings, scale parameters, or latent process variances can be hard to specify directly, while beliefs about signal-to-noise ratios or the fraction of variability explained by latent factors are often easier to articulate (Bartholomew et al., 2011).

More broadly, any modeling framework where parameter-level priors are abstract but behavioral characteristics are interpretable could benefit from reversing the direction of prior specification. The variance decomposition device developed here is specific to linear predictors, but the underlying principle has wider applicability, namely to begin with interpretable beliefs about model behavior and propagate uncertainty to parameters.

List of Figures

2.1 Posterior mean under the Horseshoe prior via Tweedie's formula 23
2.2 Shrinkage factors for the Horseshoe prior 25
2.3 Prior distribution of the effective number of nonzero coefficients 27
2.4 Implied priors on R^2 30

3.1 Beta prior for R^2 and induced Beta-Prime prior for τ^2 37
3.2 Schematic construction of the R2D2M2 prior 38
3.3 Marginal prior densities under the R2D2M2 prior 39

4.1 Implied distribution of the norm of a K-dimensional Gaussian vector 43
4.2 KL divergence between Dirichlet and Logistic-Normal distributions 44

5.1 Correlation of within-group variance allocations 47
5.2 Prior predictive distribution of group-wise effective number of nonzero coefficients 48

6.1 DASP prior contours of $p(b_1, b_2)$ for varying ρ 52

Bibliography

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). "Adapting to unknown sparsity by controlling the false discovery rate." *The Annals of Statistics*, 34(2): 584 – 653.
URL https://doi.org/10.1214/009053606000000074 16

Abril-Pla, O., Andreani, V., Carroll, C., Dong, L., Fonnesbeck, C. J., Kochurov, M., Kumar, R., Lao, J., Luhmann, C. C., Martin, O. A., Osthege, M., Vieira, R., Wiecki, T., and Zinkov, R. (2023). "PyMC: A Modern and Comprehensive Probabilistic Programming Framework in Python." *PeerJ Computer Science*, 9(e1516). 2

Agarwal, S., Dutta, S., and Mukhopadhyay, M. (2025). "Bayesian Variable Selection Under High-Dimensional Settings with Grouped Covariates." *Bayesian Analysis*.
URL https://doi.org/10.1214/25-BA1525 45

Agresti, A. and Hitchcock, D. B. (2005). "Bayesian inference for categorical data analysis." *Statistical Methods and Applications*, 14(3): 297–330.
URL https://doi.org/10.1007/s10260-005-0121-y 42

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711 – 1767.
URL https://doi.org/10.1214/23-EJS2136 xi, 41

— (2025a). "Dependency-Aware Shrinkage Priors for High Dimensional Regression."
URL https://arxiv.org/abs/2505.10715 xii

— (2025b). "Generalized Decomposition Priors on R2." *Bayesian Analysis*, 1 – 34.
URL https://doi.org/10.1214/25-BA1524 xi

Aguilar, J. E., Kohns, D., Vehtari, A., and Bürkner, P.-C. (2025). "R2 priors for Grouped Variance Decomposition in High-dimensional Regression."
URL https://arxiv.org/abs/2507.11833 xi

Aitchison, J. J. (1986). *The statistical analysis of compositional data / J. Aitchison.*. Monographs on statistics and applied probability (Series). Chapman and Hall. 41

Altman, E. I. (1968). "Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy." *The Journal of Finance*, 23(4): 589–609.
URL http://www.jstor.org/stable/2978933 16

Andrade, J. A. A. and O'Hagan, A. (2011a). "Bayesian Robustness Modelling of Location and Scale Parameters." *Scandinavian Journal of Statistics*, 38(4): 691–711.
URL http://www.jstor.org/stable/41411145 19

— (2011b). "Bayesian Robustness Modelling of Location and Scale Parameters." *Scandinavian Journal of Statistics*, 38(4): 691–711.
URL http://www.jstor.org/stable/41411145 37

Andrews, D. F. and Mallows, C. L. (1974). "Scale Mixtures of Normal Distributions." *Journal of the Royal Statistical Society. Series B (Methodological)*, 36(1): 99–102.
URL http://www.jstor.org/stable/2984774 2, 17, 22, 38

Armagan, A., Clyde, M., and Dunson, D. B. (2011). "Generalized beta mixtures of Gaussians." In *Advances in neural information processing systems*, 523–531. 2, 17, 18, 20, 26, 31, 33, 41

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018.
URL https://doi.org/10.1093/biomet/ast028 25

Baek, B., Jo, J., Kang, M., and Kim, Y. (2026). "Stochastic LASSO for extremely high-dimensional genomic data." *Scientific Reports*, 16: 5250.
URL https://doi.org/10.1038/s41598-026-35273-3 1

Bai, R. (2025). "Bayesian Group Regularization in Generalized Linear Models with a Continuous Spike-and-Slab Prior."
URL https://arxiv.org/abs/2007.07021 58

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233.
URL https://doi.org/10.1080/02331888.2019.1662017 2, 18

— (2021). "On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models." *Statistica Sinica*.
URL http://arxiv.org/abs/1807.06539 18, 20, 27, 31, 37

Banerjee, S., Castillo, I., and Ghosal, S. (2021). "Bayesian inference in high-dimensional models." ArXiv:2101.04491 [math].
URL http://arxiv.org/abs/2101.04491 2, 15, 41, 58, 59

Barndorff-Nielsen, O., Kent, J., and Sørensen, M. (1982). "Normal variance-mean mixtures and z distributions." *International Statistical Review*, 50(2): 145–159. 17, 21, 38

Barndorff-Nielsen, O. E. and Jørgensen, B. (1991). "Some parametric models on the simplex." *Journal of Multivariate Analysis*, 39(1): 106–116.
URL https://www.sciencedirect.com/science/article/pii/0047259X9190008P 41

Bartholomew, D. J., Knott, M., and Moustaki, I. (2011). *Latent Variable Models and Factor Analysis: A Unified Approach*. Wiley, 3 edition. 61

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1): 1–48.
URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01 35

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.
URL http://www.jstor.org/stable/2346101 38

Bentler, P. M. and Bonett, D. G. (1980). "Significance tests and goodness of fit in the analysis of covariance structures." *Psychological Bulletin*, 88(3): 588–606. Place: US Publisher: American Psychological Association. 45

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer. 9, 10, 11, 12, 13, 16, 27

— (2006). "The case for objective Bayesian analysis." *Bayesian Analysis*, 1(3): 385 – 402.
URL https://doi.org/10.1214/06-BA115 12

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969.
URL http://www.jstor.org/stable/26363497 3, 13, 17, 20, 24, 53

— (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." *Bayesian Analysis*, 12(4): 1105 – 1131.
URL https://doi.org/10.1214/16-BA1028 18, 24, 33, 49

— (2019). "Lasso Meets Horseshoe: A Survey." *Statistical Science*, 34(3): 405 – 427.
URL https://doi.org/10.1214/19-STS700 15, 18, 19, 21, 25, 38, 49, 52

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression." *Biometrika*, 103(4): 985–991.
URL https://doi.org/10.1093/biomet/asw042 17, 28, 46

Bhattacharya, A. and Johndrow, J. (2021). "MCMC for Global-Local Shrinkage Priors in High-Dimensional Settings." In *Handbook of Bayesian Variable Selection*, 161–178. Boca Raton: Chapman and Hall/CRC, 1 edition.
URL https://www.taylorfrancis.com/books/9781003089018/chapters/10.1201/9781003089018-7 2, 28

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2012). "Bayesian shrinkage." ArXiv:1212.6088 [math].
URL http://arxiv.org/abs/1212.6088 19

— (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490.
URL https://doi.org/10.1080/01621459.2014.960967 2, 18, 25, 33, 41

Bissiri, P. G., Holmes, C. C., and Walker, S. G. (2016). "A General Framework for Updating Belief Distributions." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5): 1103–1130. _eprint: https://academic.oup.com/jrsssb/article-

pdf/78/5/1103/49235403/jrsssb_78_5_1103.pdf.
URL https://doi.org/10.1111/rssb.12158 15

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112(518): 859–877.
URL http://dx.doi.org/10.1080/01621459.2017.1285773 10

Blei, D. M. and Lafferty, J. D. (2007). "A correlated topic model of Science." *The Annals of Applied Statistics*, 1(1): 17 – 35.
URL https://doi.org/10.1214/07-AOAS114 42

Bockting, F., Radev, S. T., and Bürkner, P.-C. (2024). "Simulation-based prior knowledge elicitation for parametric Bayesian models." *Scientific Reports*, 14(1): 17330.
URL https://www.nature.com/articles/s41598-024-68090-7 12, 13, 29, 60

Bondell, H. D. and Reich, B. J. (2012). "Consistent High-Dimensional Bayesian Variable Selection via Penalized Credible Regions." *Journal of the American Statistical Association*, 107(500): 1610–1624. PMID: 23482517.
URL https://doi.org/10.1080/01621459.2012.716344 28

Boogaart, K. and Tolosana-Delgado, R. (2013). "Analyzing Compositional Data with R." 209–253. 42

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). "Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors." *Bayesian Analysis*, 1 – 30.
URL https://doi.org/10.1214/23-BA1371 2, 4, 5, 15, 41, 45

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition. 28

Brown, P. J. and Griffin, J. E. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171 – 188.
URL https://doi.org/10.1214/10-BA507 2, 18

Buehlmann, P., Kalisch, M., and Meier, L. (2014). "High-Dimensional Statistics with a View Toward Applications in Biology." *Annual Review of Statistics and Its Application*, 1(1): 255–278.
URL https://doi.org/10.1146/annurev-statistics-022513-115545 1, 2, 5, 15, 39

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Berlin, Heidelberg: Springer.
URL https://link.springer.com/10.1007/978-3-642-20192-9 15

Bürkner, P.-C. (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1): 1–28.
URL https://www.jstatsoft.org/index.php/jss/article/view/v080i01 23, 39, 60

Bürkner, P.-C. and Charpentier, E. (2020). "Modelling monotonic effects of ordinal predictors in Bayesian regression models." *British Journal of Mathematical and Statistical Psychology*, 73(3): 420–451.
URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12195 35

Caron, A., Baio, G., and Manolopoulou, I. (2022). "Shrinkage Bayesian Causal Forests for Heterogeneous Treatment Effects Estimation." *Journal of Computational and Graphical Statistics*, 31(4): 1202–1214. 60

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, 73–80. PMLR.
URL https://proceedings.mlr.press/v5/carvalho09a.html 2, 15, 18

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.
URL http://www.jstor.org/stable/25734098 18, 19, 22, 26, 27, 33, 37

Castillo, I. (2024). "Bayesian nonparametric statistics, St-Flour lecture notes."
URL http://arxiv.org/abs/2402.16422 3, 12, 16, 22

Castillo, I., Schmidt-Hieber, J., and Vaart, A. W. v. d. (2015). "Bayesian Linear Regression with Sparse Priors." *The Annals of Statistics*, 43(5): 1986–2018.
URL https://www.jstor.org/stable/43818568 16, 33

Castillo, I. and Vaart, A. W. v. d. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069 – 2101.
URL https://doi.org/10.1214/12-AOS1029 2, 15

Catalina, A., Bürkner, P.-C., and Vehtari, A. (2022). "Projection Predictive Inference for Generalized Linear and Additive Multilevel Models." In Camps-Valls, G., Ruiz, F. J. R., and Valera, I. (eds.), *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, 4446–4461. PMLR.
URL https://proceedings.mlr.press/v151/catalina22a.html 28

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). "Scalable Inference for Logistic-Normal Topic Models." In Burges, C. J., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc.
URL https://proceedings.neurips.cc/paper_files/paper/2013/file/285f89b802bcb2651801455c86d78f2a-Paper.pdf 42

Christensen, R. (2020). *Plane Answers to Complex Questions: The Theory of Linear Models*. Springer Texts in Statistics. Cham: Springer International Publishing.
URL http://link.springer.com/10.1007/978-3-030-32097-3 5, 15

Cressie, N. and Wikle, C. K. (2011). *Statistics for Spatio-Temporal Data*. Wiley. 61

Datta, J. and Ghosh, J. K. (2013). "Asymptotic Properties of Bayes Risk for the Horseshoe Prior." *Bayesian Analysis*, 8(1): 111–132.
URL https://projecteuclid.org/journals/bayesian-analysis/volume-8/issue-1/Asymptotic-Properties-of-Bayes-Risk-for-the-Horseshoe-Prior/10.1214/13-BA805.full 22, 27

Datta, J., Polson, N., and Sokolov, V. (2025). "Bayesian Global-Local Regularization." ArXiv:2512.13992 [stat].
URL http://arxiv.org/abs/2512.13992 16

Dawid, A. P., Stone, M., and Zidek, J. V. (1973). "Marginalization Paradoxes in Bayesian and Structural Inference." *Journal of the Royal Statistical Society. Series B (Methodological)*, 35(2): 189–233.
URL http://www.jstor.org/stable/2984907 19

DeGroot, M. H. (1970). *Optimal Statistical Decisions*. New York: McGraw-Hill. 10

Efron, B. (1978). "Controversies in the Foundations of Statistics." *The American Mathematical Monthly*, 85(4): 231–246.
URL http://www.jstor.org/stable/2321163 3, 23, 53

— (2011). "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association*, 106(496): 1602–1614.
URL https://doi.org/10.1198/jasa.2011.tm11181 16, 22

Fazio, L., Scholz, M., Aguilar, J. E., and Bürkner, P.-C. (2025). "Primed Priors for Simulation-Based Validation of Bayesian Models."
URL https://arxiv.org/abs/2408.06504 xii

Fuglstad, G.-A., Hem, I. G., Knight, A., Rue, H., and Riebler, A. (2020). "Intuitive Joint Priors for Variance Parameters." *Bayesian Analysis*, 15(4): 1109 – 1137.
URL https://doi.org/10.1214/19-BA1185 4, 13

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). "Constructing Priors that Penalize the Complexity of Gaussian Random Fields." *Journal of the American Statistical Association*, 114(525): 445–452.
URL https://doi.org/10.1080/01621459.2017.1415907 4, 35

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). "Visualization in Bayesian Workflow." *Journal of the Royal Statistical Society Series A: Statistics in Society*, 182(2): 389–402.
URL http://dx.doi.org/10.1111/rssa.12378 13

Ge, H., Xu, K., and Ghahramani, Z. (2018). "Turing: A Language for Flexible Probabilistic Inference." In Storkey, A. and Perez-Cruz, F. (eds.), *Proceedings of the Twenty-First*

International Conference on Artificial Intelligence and Statistics, volume 84 of *Proceedings of Machine Learning Research*, 1682–1690. PMLR.
URL https://proceedings.mlr.press/v84/ge18b.html 2

Gelfand, A. E. and Smith, A. F. M. (1990). "Sampling-Based Approaches to Calculating Marginal Densities." *Journal of the American Statistical Association*, 85(410): 398–409.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10476213 10

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515 – 534.
URL https://doi.org/10.1214/06-BA117A 3, 19, 23

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 1, 9, 10, 13, 35, 58

Gelman, A., Goodrich, B., Gabry, J., and Vehtari, A. (2019). "R-squared for Bayesian Regression Models." *The American Statistician*, 73(3): 307–309.
URL https://doi.org/10.1080/00031305.2018.1549100 29

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press. 1, 4, 35, 36

Gelman, A., Hill, J., and Vehtari, A. (2020a). *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press. 1, 4, 12, 13, 29

Gelman, A., Simpson, D., and Betancourt, M. (2017). "The Prior Can Often Only Be Understood in the Context of the Likelihood." *Entropy*, 19(10): 555.
URL http://dx.doi.org/10.3390/e19100555 3, 12

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020b). "Bayesian Workflow."
URL https://arxiv.org/abs/2011.01808 3, 10, 11, 12, 13, 14, 17

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889.
URL https://www.jstor.org/stable/2290777 17, 19, 49

— (1995). "Stochastic search variable selection." *Markov Chain Monte Carlo in Practice*. 49

— (1997). "Approaches for Bayesian Variable Selection." *Statistica Sinica*, 7(2): 339–373.
URL https://www.jstor.org/stable/24306083 17, 19

Ghosal, S., Ghosh, J. K., and Vaart, A. W. v. d. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500 – 531.
URL https://doi.org/10.1214/aos/1016218228 12

Ghosal, S. and Vaart, A. W. v. d. (2017). *Fundamentals of Nonparametric Bayesian Inference*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 12

Ghosh, J. K. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics Ser. New York, NY: Springer New York. 2, 12, 15

Ghosh, P. and Chakrabarti, A. (2017). "Asymptotic Optimality of One-Group Shrinkage Priors in Sparse High-dimensional Problems." *Bayesian Analysis*, 12(4): 1133 – 1161.
URL https://doi.org/10.1214/16-BA1029 2, 20, 21, 25

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). "Model Selection in Bayesian Neural Networks via Horseshoe Priors." *Journal of Machine Learning Research*, 20(182): 1–46.
URL http://jmlr.org/papers/v20/19-236.html 58

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. 2, 4, 15

Gneiting, T. and Raftery, A. E. (2007). "Strictly Proper Scoring Rules, Prediction, and Estimation." *Journal of the American Statistical Association*, 102(477): 359–378.
URL https://doi.org/10.1198/016214506000001437 14

Goldstein, M. (2006). "Subjective Bayesian Analysis: Principles and Practice." *Bayesian Analysis*, 1(3): 403 – 420.
URL https://doi.org/10.1214/06-BA116 12

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). "rstanarm: Bayesian applied regression modeling via Stan."
URL https://mc-stan.org/rstanarm 23

Gorinova, M. I., Gordon, A. D., and Sutton, C. (2019). "Probabilistic Programming with Densities in SlicStan: Efficient, Flexible, and Deterministic." *Proc. ACM Program. Lang.*, 3(POPL).
URL https://doi.org/10.1145/3290348 29

Goutis, C. and Robert, C. P. (1998). "Model Choice in Generalised Linear Models: A Bayesian Approach Via Kullback-Leibler Projections." *Biometrika*, 85(1): 29–37.
URL http://www.jstor.org/stable/2337307 28

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., and Quinn, T. (2023). "Aitchison's Compositional Data Analysis 40 Years on: A Reappraisal." *Statistical Science*, 38(3): 386 – 410.
URL https://doi.org/10.1214/22-STS880 42

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. (2012). "A Kernel Two-Sample Test." *Journal of Machine Learning Research*, 13(25): 723–773.
URL http://jmlr.org/papers/v13/gretton12a.html 60

Griffin, J. and Brown, P. (2017). "Hierarchical Shrinkage Priors for Regression Models." *Bayesian Analysis*, 12(1): 135 – 159.
URL https://doi.org/10.1214/15-BA990 4, 59

Griffin, J. E. and Brown, P. J. (2013). "Some Priors for Sparse Regression Modelling." *Bayesian Analysis*, 8(3): 691 – 702.
URL https://doi.org/10.1214/13-BA827 2

— (2021). "Bayesian global-local shrinkage methods for regularisation in the high dimension linear model." *Chemometrics and Intelligent Laboratory Systems*, 210: 104255.
URL https://www.sciencedirect.com/science/article/pii/S016974392100023X 4

Griffin, M. and Hoff, P. D. (2024). "Structured Shrinkage Priors." *Journal of Computational and Graphical Statistics*, 33(1): 1–14.
URL https://www.tandfonline.com/doi/citedby/10.1080/10618600.2023.2233577 4, 49

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association*, 110(509): 435–448.
URL https://doi.org/10.1080/01621459.2014.993077 28, 60

Hardin, J. and Hilbe, J. (2007). *Generalized Linear Models and Extensions, Second Edition*. A Stata Press publication. Taylor & Francis.
URL https://books.google.de/books?id=tOeqO6Hs-6gC 29, 57

Harrell, F. (2013). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer New York. 1, 4, 35

Hartmann, M., Agiashvili, G., Bürkner, P., and Klami, A. (2020). "Flexible Prior Elicitation via the Prior Predictive Distribution."
URL https://arxiv.org/abs/2002.09868 12, 13

Hastie, T. and Tibshirani, R. (1986). "Generalized Additive Models." *Statistical Science*, 1(3): 297–310.
URL http://www.jstor.org/stable/2245459 4

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Science & Business Media. 15

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. ISSN. CRC Press. 15

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67.
URL http://www.jstor.org/stable/1267351 15, 16

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15(1): 1593–1623. 2, 28

Huang, J., Breheny, P., and Ma, S. (2012). "A Selective Review of Group Selection in High-Dimensional Models." *Statistical Science*, 27(4): 481–499. 45

Iorio, M. D., Heinecke, A., Franzolini, B., and Cabral, R. (2025). "Singularities in Bayesian Inference: Crucial or Overstated?"
URL https://arxiv.org/abs/2501.06618 18, 21, 22, 25, 55

Izmailov, P., Vikram, S., Hoffman, M. D., and Wilson, A. G. (2021). "What Are Bayesian Neural Network Posteriors Really Like?"
URL https://arxiv.org/abs/2104.14421 58

Jeong, S. and Ghosal, S. (2020). "Posterior contraction in sparse generalized linear models." *Biometrika*, 108(2): 367–379.
URL https://doi.org/10.1093/biomet/asaa074 58

Jessen, A. H. and Mikosch, T. (2006). "Regularly Varying Functions." *Publications de l'Institut Mathématique*, 80(94)(100): 171–192.
URL https://eudml.org/doc/258304 20, 22

Jiang, W. (2007). "Bayesian variable selection for high dimensional generalized linear models: Convergence rates of the fitted densities." *The Annals of Statistics*, 35(4).
URL http://dx.doi.org/10.1214/009053607000000019 58

Joachims, T. (1998). "Text Categorization with Support Vector Machines: Learning with Many Relevant Features." In *Machine Learning: ECML-98*, volume 1398 of *Lecture Notes in Computer Science*. Springer. 1

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *The Annals of Statistics*, 32(4): 1594 – 1649.
URL https://doi.org/10.1214/009053604000000030 15

Kang, J., Bang, S., and Jhun, M. (2016). "Hierarchically penalized quantile regression." *Journal of Statistical Computation and Simulation*, 86(2): 340–356. 45

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer, 2 edition. 61

Kohns, D. and Potjagailo, G. (2025). "Flexible Bayesian MIDAS: time-variation, group-shrinkage and sparsity." *Journal of Business & Economic Statistics*, 1–17. 45

Kruijer, W., Rousseau, J., and Vaart, A. W. v. d. (2010). "Adaptive Bayesian Density Estimation with Location-Scale Mixtures." *Electronic Journal of Statistics*, 4. 32

Kruschke, J. K. (2015). "Chapter 6 - Inferring a Binomial Probability via Exact Mathematical Analysis." In Kruschke, J. K. (ed.), *Doing Bayesian Data Analysis (Second Edition)*, 123–141. Boston: Academic Press, second edition edition.
URL https://www.sciencedirect.com/science/article/pii/B9780124058880000064 31

Ledoit, O. and Wolf, M. (2004). "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis*, 88(2): 365–411.
URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964 51

— (2020). "Analytical Nonlinear Shrinkage of Large-Dimensional Covariance Matrices." *The Annals of Statistics*, 48(5): 3043–3065.
URL https://www.jstor.org/stable/27028732 51

Lee, K., Jo, S., and Lee, J. (2022). "The beta-mixture shrinkage prior for sparse covariances with near-minimax posterior convergence rate." *Journal of Multivariate Analysis*, 192: 105067. 59

Lehmann, E. and Casella, G. (2006). *Theory of Point Estimation*. Springer Texts in Statistics. Springer New York.
URL https://books.google.de/books?id=4f24CgAAQBAJ 16

Li, Y., Craig, B. A., and Bhadra, A. (2019). "The graphical horseshoe estimator for inverse covariance matrices." *Journal of Computational and Graphical Statistics*, 28(3): 747–757. 59

Lin, J. (2016). "On The Dirichlet Distribution by Jiayu Lin." 32

Lin, L., Drton, M., and Shojaie, A. (2020). "Statistical Significance in High-Dimensional Linear Mixed Models." In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS '20, 171–181. New York, NY, USA: Association for Computing Machinery.
URL https://doi.org/10.1145/3412815.3416883 4, 39

Lindgren, F. and Rue, H. (2015). "Bayesian Spatial Modelling with R-INLA." *Journal of Statistical Software*, 63(19): 1–25. 1

Lindgren, F., Rue, H., and Lindström, J. (2011). "An Explicit Link between Gaussian Fields and Gaussian Markov Random Fields: The Stochastic Partial Differential Equation Approach." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(4): 423–498. 1

Lindley, D. V. (1972). *Bayesian Statistics: A Review*. Philadelphia, PA: SIAM. 9, 10

Linero, A. R. (2017). "A review of tree-based Bayesian methods." *Communications for Statistical Applications and Methods*, 24(6): 543–559.
URL http://www.csam.or.kr/journal/view.html?doi=10.29220/CSAM.2017.24.6.543 60

— (2023). "Prior and Posterior Checking of Implicit Causal Assumptions." *Biometrics*, 79(4): 3153–3164. 3, 60

— (2024). "In Nonparametric and High-Dimensional Models, Bayesian Ignorability is an Informative Prior." *Journal of the American Statistical Association*, 119(548): 2785–2798. 60

Makalic, E. and Schmidt, D. F. (2016). "A Simple Sampler for the Horseshoe Estimator." *IEEE Signal Processing Letters*, 23(1): 179–182.
URL http://dx.doi.org/10.1109/LSP.2015.2503725 17, 28, 46

Malone, B. M., Tan, F., Bridges, S. M., and Peng, Z. (2011). "Comparison of Four ChIP-Seq Analytical Algorithms Using Rice Endosperm H3K27 Trimethylation Profiling Data." *PLOS ONE*, 6(9): 1–12.
URL https://doi.org/10.1371/journal.pone.0025260 1, 15

Manderson, A. A. and Goudie, R. J. B. (2026). "Translating predictive distributions into informative priors." *Journal of the American Statistical Association*, 0(ja): 1–18.
URL https://doi.org/10.1080/01621459.2026.2614034 3, 11, 12, 13, 60

Marčenko, V. A. and Pastur, L. A. (1967). "Distribution of eigenvalues for some sets of random matrices." *Mathematics of the USSR-Sbornik*, 1(4): 457. 15, 51

Masreliez, C. (1975). "Approximate non-Gaussian filtering with linear state and observation relations." *IEEE Transactions on Automatic Control*, 20(1): 107–110. 22

Matabuena, M., Ghosal, R., Aguilar, J. E., Keshet, A., Wagner, R., Fernández Merino, C., Sánchez Castro, J., Zipunnikov, V., Onnela, J.-P., and Gude, F. (2025). "Glucodensity functional profiles outperform traditional continuous glucose monitoring metrics." *Scientific Reports*, 15: 33662. xii

McLatchie, Y., Rögnvaldsson, S., Weber, F., and Vehtari, A. (2025). "Advances in Projection Predictive Inference." *Statistical Science*, 40(1): 128 – 147.
URL https://doi.org/10.1214/24-STS949 28

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). "Prior knowledge elicitation: The past, present, and future." 3, 12

Miller, P. (2006). *Applied Asymptotic Analysis*. American Mathematical Society, 1 edition. 20

Mimno, D., Wallach, H., and McCallum, A. (2008). "Gibbs sampling for logistic normal topic models with graph-based priors." 42

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, 83(404): 1023–1032.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694 17

Mogliani, M. and Simoni, A. (2021). "Bayesian MIDAS penalized regressions: Estimation, selection, and prediction." *Journal of Econometrics*, 222(1): 833–860.
URL https://linkinghub.elsevier.com/retrieve/pii/S0304407620302207 45

Morris, M., Wheeler-Martin, K., Simpson, D. P., Mooney, S. J., Gelman, A., and DiMaggio, C. J. (2019). "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan." *Spatial and spatio-temporal epidemiology*, 31: 100301. 35

Mukherjee, S., Aguilar, J. E., Zago, M., Claassen, M., and Bürkner, P.-C. (2025). "Latent variable estimation with composite Hilbert space Gaussian processes."
URL https://arxiv.org/abs/2510.25371 xii

Murphy, K. P. (2022). *Probabilistic Machine Learning: An introduction*. MIT Press.
URL http://probml.github.io/book1 15, 60

Nakagawa, S. and Schielzeth, H. (2013). "A general and simple method for obtaining R2 from generalized linear mixed-effects models." *Methods in Ecology and Evolution*, 4(2): 133–142.
URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210x.2012.00261.x 29, 57

Neal, R. M. (2011). *MCMC using Hamiltonian dynamics*.
URL http://arxiv.org/abs/1206.1901 2, 28

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer, 2 edition. 58

O'Hagan, A. and Pericchi, L. (2012). "Bayesian heavy-tailed models and conflict resolution: A review." *Brazilian Journal of Probability and Statistics*, 26(4): 372–401.
URL https://www.jstor.org/stable/43601225 19

Olver, F., Standards, N. I. o., (U.S.), T., Lozier, D., Boisvert, R., and Clark, C. (2010). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press. 20

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686.
URL https://doi.org/10.1198/016214508000000337 15, 19

Pas, S. L. v. d. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC. 2, 3, 7, 15, 18, 21, 26, 27, 38

Pas, S. L. v. d., Kleijn, B. J. K., and Vaart, A. W. v. d. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585 – 2618.
URL https://doi.org/10.1214/14-EJS962 2, 20, 21

Pas, S. L. v. d., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1):

976 – 1000.
URL https://doi.org/10.1214/16-EJS1130 2, 17, 19, 20, 21, 22, 25

Pas, S. L. v. d., Szabó, B., and Vaart, A. W. v. d. (2017a). "Adaptive posterior contraction rates for the horseshoe." *Electronic Journal of Statistics*, 11(2).
URL https://projecteuclid.org/journals/electronic-journal-of-statistics/volume-11/issue-2/Adaptive-posterior-contraction-rates-for-the-horseshoe/10.1214/17-EJS1316.full 18, 25

— (2017b). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221 – 1274.
URL https://doi.org/10.1214/17-BA1065 17, 27, 47

Pauger, D. and Wagner, H. (2019). "Bayesian Effect Fusion for Categorical Predictors." *Bayesian Analysis*, 14(2): 341 – 369.
URL https://doi.org/10.1214/18-BA1096 45

Pawlowsky-Glahn, V. and Buccianti, A. (2011). "Compositional data analysis : theory and applications."
URL https://api.semanticscholar.org/CorpusID:117956959 41

Pericchi, L. R. and Smith, A. F. M. (1992). "Exact and Approximate Posterior Moments for a Normal Location Parameter." *Journal of the Royal Statistical Society. Series B (Methodological)*, 54(3): 793–804.
URL http://www.jstor.org/stable/2345859 22

Piironen, J., Paasiniemi, M., Catalina, A., Weber, F., and Vehtari, A. (2023). "projpred: Projection Predictive Feature Selection."
URL https://mc-stan.org/projpred/ 28

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155 – 2197.
URL https://doi.org/10.1214/20-EJS1711 5, 27, 28, 49

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018 – 5051.
URL https://doi.org/10.1214/17-EJS1337SI 2, 7, 17, 18, 19, 26, 38, 47

Polson, N. G. (1991). "A Representation of the Posterior Mean for a Location Model." *Biometrika*, 78(2): 426–430.
URL http://www.jstor.org/stable/2337270 22

Polson, N. G. and Scott, J. G. (2011). "Shrink Globally, Act Locally: Sparse Bayesian Regularization and Prediction*." In Bernardo, J. M., Bayarri, M. J., Berger, J. O., Dawid, A. P., Heckerman, D., Smith, A. F. M., and West, M. (eds.), *Bayesian Statistics 9*, 501–538. Oxford University Press.

URL https://academic.oup.com/book/1879/chapter/141655378 2, 7, 17, 18, 19, 21, 22, 49

— (2012a). "Local Shrinkage Rules, Lévy Processes and Regularized Regression." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(2): 287–311.
URL https://doi.org/10.1111/j.1467-9868.2011.01015.x 21

— (2012b). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7(4): 887 – 902.
URL https://doi.org/10.1214/12-BA730 17, 18, 19, 31

Quackenbush, J. (2002). "Microarray data normalization and transformation." *Nature Genetics*, 32(S4): 496–501.
URL https://www.nature.com/articles/ng1032z 16

Rasmussen, C. E. and Williams, C. K. I. (2006). *Gaussian Processes for Machine Learning*. The MIT Press. 58

Reiser, P., Aguilar, J. E., Guthke, A., and Bürkner, P.-C. (2025). "Uncertainty quantification and propagation in surrogate-based Bayesian inference." *Statistics and Computing*, 35: 66. xii

Rights, J. D. and Sterba, S. K. (2019). "Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures." *Psychological methods*, 24 3: 309–338. 29, 57

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York. 9, 10, 17, 23, 27, 38

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer.
URL http://link.springer.com/10.1007/978-1-4757-4145-2 10

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444.
URL https://doi.org/10.1080/01621459.2016.1260469 33

Rényi, A. (1961). "On measures of entropy and information." In *Proceedings of the fourth Berkeley symposium on mathematical statistics and probability, volume 1: contributions to the theory of statistics*, volume 4, 547–562. University of California Press. 60

Savage, L. J. (1954). *The Foundations of Statistics*. New York: John Wiley & Sons. 10

Seneta, E. (1976). *Regularly Varying Functions*, volume 508 of *Lecture Notes in Mathematics*. Berlin, Heidelberg: Springer-Verlag. 20

Soares, J. M., Magalhães, R., Moreira, P. S., Sousa, A., Ganz, E., Sampaio, A., Alves, V., Marques, P., and Sousa, N. (2016). "A Hitchhiker's Guide to Functional Magnetic Resonance Imaging." *Frontiers in Neuroscience*, 10: 515. 1

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high-dimensional regression." *Science China Mathematics*, 66: 409–442.
URL https://api.semanticscholar.org/CorpusID:88516067 12, 22, 25, 33

Stan Development Team (2026). "Stan Modeling Language Users Guide and Reference Manual, Version 2.38."
URL http://mc-stan.org/ 2, 7, 29, 39

Stein, C. (1956). "Inadmissibility of the usual estimator for the mean of a multivariate normal distribution." In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, volume 3, 197–207. University of California Press. 16

Stein, C. M. (1981). "Estimation of the Mean of a Multivariate Normal Distribution." *The Annals of Statistics*, 9(6): 1135–1151.
URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6/Estimation-of-the-Mean-of-a-Multivariate-Normal-Distribution/10.1214/aos/1176345632.full 16

Sun, H., Lin, W., Feng, R., and Li, H. (2014). "Network-regularized high-dimensional Cox regression for analysis of genomic data." *Statistica Sinica*, 24(3): 1433–1459. 45

Tadesse, M. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. 2, 18, 19, 24, 27, 46, 55

Takada, Y. (1979). "Stein's positive part estimator and bayes estimator." *Annals of the Institute of Statistical Mathematics*, 31(2): 177–183.
URL http://link.springer.com/10.1007/BF02480275 16

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). "Validating Bayesian Inference Algorithms with Simulation-Based Calibration." 38

Tang, Z., Shen, Y., Zhang, X., and Yi, N. (2017). "The Spike-and-Slab Lasso Generalized Linear Models for Prediction and Associated Genes Detection." *Genetics*, 205(1): 77–88.
URL https://doi.org/10.1534/genetics.116.192195 58

Tew, S. Y., Schmidt, D. F., and Boley, M. (2025). "GRASP: Grouped Regression with Adaptive Shrinkage Priors."
URL https://arxiv.org/abs/2506.18092 45

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. 15, 16

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., Vehtari, A., and Hsu, D. (2022). "The piranha problem: Large effects swimming in a small pond." 15

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2023). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." URL https://mc-stan.org/loo/ 14, 28, 38

Vehtari, A., Gelman, A., and Gabry, J. (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. URL http://dx.doi.org/10.1007/s11222-016-9696-4 13, 14

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667 – 718. URL https://doi.org/10.1214/20-BA1221 28

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142 – 228. URL https://doi.org/10.1214/12-SS102 13, 14

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press. 24, 42

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. 1, 15

Wasserman, L. (2004). *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. New York, NY: Springer New York. URL http://link.springer.com/10.1007/978-0-387-21736-9 1, 16

West, M. (1984). "Outlier Models and Prior Distributions in Bayesian Linear Regression." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 46(3): 431–439. URL https://academic.oup.com/jrsssb/article/46/3/431/7028219 19

— (1987). "On scale mixtures of normal distributions." *Biometrika*, 74(3): 646–648. URL https://doi.org/10.1093/biomet/74.3.646 2, 17, 38, 41

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall / CRC texts in statistical science. CRC Press/Taylor & Francis Group. 35

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). "Bayesian Grouped Horseshoe Regression with Application to Additive Models." In Kang, B. H. and Bai, Q. (eds.), *AI 2016: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, 229–240. Cham: Springer International Publishing. 5, 45

Zanella, G. and Roberts, G. (2021). "Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions (with Discussion)." *Bayesian Analysis*, 16(4): 1309 – 2770.
URL https://doi.org/10.1214/20-BA1242 29

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques*. 50

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823 – 844.
URL https://doi.org/10.1214/17-BA1076 13, 27, 28, 29, 47

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13.
URL https://doi.org/10.1080/01621459.2020.1825449 2, 3, 5, 13, 18, 20, 30, 31, 32, 33, 38, 41

Zhang, Z., Wang, S., Liu, D., et al. (2012). "EP-GIG Priors and Applications in Bayesian Sparse Learning." *Journal of Machine Learning Research*, 13(6). 18

Zhou, X., Heng, Q., Chi, E. C., and Zhou, H. (2024). "Proximal MCMC for Bayesian inference of constrained and regularized estimation." *The American Statistician*, 78(4): 379–390. 59

Publications

Article I

Aguilar, J. E., and Bürkner, P.-C. (2023). Intuitive joint priors for Bayesian linear multilevel models: the R2D2M2 prior. *Electronic Journal of Statistics* 17(1): 1711–1767.

DOI: https://doi.org/10.1214/23-EJS2136

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Electronic Journal of Statistics
Vol. 17 (2023) 1711–1767
ISSN: 1935-7524
https://doi.org/10.1214/23-EJS2136

Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior

Javier Enrique Aguilar[1] **and Paul-Christian Bürkner**[2]

[1] *Cluster of Excellence SimTech, University of Stuttgart*
e-mail: javier.aguilar-romero@simtech.uni-stuttgart.de
url: https://jear2412.github.io

[2] *Cluster of Excellence SimTech, University of Stuttgart*
e-mail: paul.buerkner@gmail.com
url: https://paul-buerkner.github.io

Abstract: The training of high-dimensional regression models on comparably sparse data is an important yet complicated topic, especially when there are many more model parameters than observations in the data. From a Bayesian perspective, inference in such cases can be achieved with the help of shrinkage prior distributions, at least for generalized linear models. However, real-world data usually possess multilevel structures, such as repeated measurements or natural groupings of individuals, which existing shrinkage priors are not built to deal with.

We generalize and extend one of these priors, the R2D2 prior by Zhang et al. (2020), to linear multilevel models leading to what we call the R2D2M2 prior. The proposed prior enables both local and global shrinkage of the model parameters. It comes with interpretable hyperparameters, which we show to be intrinsically related to vital properties of the prior, such as rates of concentration around the origin, tail behavior, and amount of shrinkage the prior exerts.

We offer guidelines on how to select the prior's hyperparameters by deriving shrinkage factors and measuring the effective number of non-zero model coefficients. Hence, the user can readily evaluate and interpret the amount of shrinkage implied by a specific choice of hyperparameters.

Finally, we perform extensive experiments on simulated and real data, showing that our inference procedure for the prior is well calibrated, has desirable global and local regularization properties and enables the reliable and interpretable estimation of much more complex Bayesian multilevel models than was previously possible.

Keywords and phrases: Bayesian inference, multilevel models, prior specification, shrinkage priors, regularization.

Received August 2022.

1. Introduction

Regression models are ubiquitous in the quantitative sciences and industry, making up a big part of all statistical data analyses. Their success can be explained by a combination of several factors, involving the ease of interpretation of their additive structure, rich mathematical theory, and (relative) simplicity of their estimation, to name a few key aspects (Gelman et al., 2020a; Harrell, 2013).

Despite these advantages, the vast amount of modeling options and required analyst choices render building interpretable, robust, and well-predicting regression models a highly difficult task.

As the complexity of the analyzed data increases, so does the required complexity of the applied regression models and we are bound to encounter multilevel structures which contain overall (population-level or fixed effects) and varying (group-specific, group-level, or random effects) terms. Overall effects influence every member of the population equally, whereas varying effects do so per group. For example, multilevel structures can be found in Psychology, Medicine, or Biology, due to the natural groupings of individuals, or repeated measurement of the same individuals in experiments or longitudinal studies. Multilevel models are designed specifically to account for the nested structure in multilevel data and are a widely applied class of regression models (Bates et al., 2015; Gelman and Hill, 2006; Bürkner, 2021). Their key statistical idea is to assume that a set of regression coefficients, defined according to the multilevel structure, originate from the same underlying distribution whose hyperparameters are subsequently estimated from the data. The implied partial-pooling property increases the robustness of the parameter estimates, helps with uncertainty calibration, and improves out-of- sample predictive performance (Gelman et al., 2013; Gelman and Hill, 2006). While multilevel models can be estimated in both frequentist and Bayesian frameworks (Bates et al., 2015; Bürkner, 2017), we will focus on the latter framework in this work, since it offers considerable flexibility in the specification and regularization of multilevel models (Gelman et al., 2013).

From a Bayesian perspective, the above mentioned distributions of the regression coefficients constitute prior distributions (priors) that describe the uncertainty in the model parameters before seeing the data. The desirable properties of multilevel models can then be explained by the fact that these form joint priors over a set of parameters with shared hyperparameters, rather than separate independent priors for each parameter (Gelman et al., 2013). Joint priors are a Bayesian success story more generally. For example, joint priors can help to improve the predictive performance of individual additive terms parameterized by more than one parameter, for example, splines, spatial, or monotonic effects (Bürkner and Charpentier, 2020; Wood, 2017; Morris et al., 2019). With very few exceptions (Fuglstad et al., 2019; Goodrich et al., 2020; Yanchenko et al., 2021), the development of joint priors for Bayesian multilevel models has been limited to individual additive terms. In contrast, different terms, corresponding to different parameter sets, still receive mutually independent priors, for example, independent (inverse-)gamma, uniform, or half Student-t priors for variance or standard deviation parameters (Browne and Draper, 2006; Bürkner, 2017; Simpson et al., 2017; Depaoli and Clifton, 2015). As more and more terms are being added to the model while the number of observations remains constant, such models will overfit the data, leading to unreliable or uninterpretable estimates as well as bad out-of-sample predictions (Van Erp et al., 2019). Also, overall Type I error rates may be severely inflated if many terms with mutually independent priors are tested (Gelman et al., 2013).

Without advanced regularization methods such as joint priors, reliable and interpretable estimation of the required highly parameterized models is very hard to achieve. This is particularly obvious in the high dimensional case regime (more covariates than observations: $p > N$), where the number of model parameters p becomes larger than the number of observations N in the data and unregularized regression models would not be estimable (Hoerl and Kennard, 1970); but even if $p < N$, regression models often benefit strongly from regularization, for example, to prevent overfitting and Type I error inflation (Hoerl and Kennard, 1970; Gelman et al., 2013; Van Erp et al., 2019). While all of these difficulties apply to standard regression models already, they become even worse for multilevel models as the number of additive terms to explore further increases when considering all valid combinations of predictor coefficients and grouping factors which are permitted by the given data structure (Catalina et al., 2020; Barr et al., 2013; Paananen et al., 2020).

In the context of *single-level* linear models, that is, linear models without multilevel structure, a wide range of joint shrinkage priors have been developed (Park and Casella, 2008; Carvalho et al., 2010; Piironen and Vehtari, 2017; Bhattacharya et al., 2015). These models enable the estimation, interpretation, and selection of predictor terms and prevent overfitting even when the number of predictors is large and/or the data is small (Van Erp et al., 2019; van der Pas, 2021). On an abstract level, these priors can all be described as Global-Local (GL) priors because they have both, global parameters controlling the shrinkage of all terms jointly and local parameters controlling the (relative) shrinkage of individual terms (van der Pas, 2021). Usually, they are combined with a location-scale distribution, such as normal or a double exponential, where the location is fixed to zero and the scale is computed as a product of global and local parameters. While these priors have been very successful within their supported model class (Van Erp et al., 2019; van der Pas, 2021), they remain limited in several key aspects which we aim to address in our work:

1. Most existing joint priors for regression models require the manual specification of hyperparameters with potentially strong impact on the obtained inference (Van Erp et al., 2019). These hyperparameters are often unintuitive and abstract especially for non-statistical experts, making these priors' practical application and communication much more difficult.
2. High dimensionality in multilevel models has so far been studied in cases where the number of overall coefficients p is big while the amount of varying terms q is very small, $p > N, q \ll N$. For example, in their papers, Buehlmann et al. (2014) and Lin et al. (2020) carried out theoretical research and simulation studies of frequentist multilevel models with large p yet $q \leq 2$.
3. Existing joint priors on multilevel models focus only on the group-specific coefficients (Fuglstad et al., 2019; Goodrich et al., 2020), without at the same time considering the overall coefficients. Further, it remains unclear how these priors scale with the number of predictors when their coefficients vary across an increasing number of grouping factors. However, in order to

ensure consistent and robust regularization in multilevel models, we need to define a single global joint prior on all coefficients, and ensure it scales well to large numbers of additive terms and complex multilevel structure.

An initial step towards addressing these issues, at least for single-level linear models, is the R2D2 prior developed by Zhang et al. (2020). The main idea is to specify a prior on the coefficient of determination R^2 – also known as *proportion of explained variance* – and decompose the *explained variance* into individual variance components via Dirichlet Decomposition (abbreviated as D2), thus propagating uncertainty from R^2 on to the regression coefficients. We further motivate the use of priors over R^2 in Section 2. The R2D2 prior shares several desirable properties with other global-local shrinkage priors for single-level linear models (Bhadra et al., 2016; Carvalho et al., 2010; Zhang et al., 2020) and is easy to understand even for non-expert analysts since the only hyperparameters it requires are the shape parameters of the prior on R^2 and a single concentration parameter for the Dirichlet distribution. However, its scope is limited to single-level linear models. These simple models do not live up to the requirements of model-based research, where multilevel data structures are omnipresent, so a much more general solution is needed.

Therefore, the aim of our paper is to fill this gap by generalizing the R2D2 prior to models with increased complexity in terms of multilevel structures. We name our new prior the *R2D2M2 prior*, where M2 stands for *M*ultilevel *M*odels. In parallel to our work, related ideas were put forward by Yanchenko et al. (2021) although with a different focus: they concentrate on defining useful R^2 measures for non normal likelihoods without studying the theoretical properties on the resulting priors as we do here.

1.1. Main contributions

Our main contributions are as follows:

- We propose a prior over a global R^2 that jointly regularizes both the overall and varying coefficients at the same time. Our new prior's hyperparameters are easier to interpret, which facilitates their specification by the user.
- We expand upon the theoretical properties of the original R2D2 prior by considering normal base for distributions for the coefficients. We study concentration properties around the origin and behavior in the tails and show how both are dependent on the hyperparameters selected by the user.
- We propose the use of shrinkage factors to evaluate how the chosen hyperparameters determine the amount of shrinkage for each individual coefficient. We also propose a global measure of shrinkage, which quantifies the effective total amount of non-zero coefficients in the model. Both of these quantities were not considered for the R2D2 prior before.
- We implement both the R2D2 prior and the R2D2M2 prior in the probabilistic programming language Stan (Stan Development Team, 2022; Car-

penter et al., 2017) and demonstrate that inference for our implementations is well calibrated. We also provide an implementation in the R package `brms` (Bürkner, 2017, version 2.19.2 or higher), which provides a high level interface to fit Bayesian generalized (non-)linear multivariate multilevel models using Stan. We further perform extensive simulations to test the capabilities of the R2D2M2 prior and discuss its shrinkage behavior, out-of-sample predictive performance, as well as the relationship of the latter with the amount of shrinkage. We also show that frequentist error metrics such as Type I and Type II errors can be controlled with the help of our prior even in complex scenarios.

- Our joint prior offers the possibility to study cases in which both the number of overall coefficients p and varying terms q are large and potentially both greater than N, that is, high dimensionality in both the overall and varying coefficients. It is to the best of our knowledge the first method that provides this. To demonstrate the prior's applicability and usefulness in such high-dimensional cases not only in simulations but also in the real world, we reanalyse the riboflavin dataset provided by Buehlmann et al. (2014).

The remainder of the paper is organized as follows: Section 2 starts by motivating the use of priors over R^2 (and thus the use of the R2D2M2 prior) by showing how weakly informative priors on the coefficients can translate into very informative priors for R^2. The construction of the R2D2M2 prior is presented along with insights that lead to better understanding of the prior. Section 3 presents theoretical properties of both the R2D2 and R2D2M2 prior such as marginal distributions and concentration properties as well as guidelines for hyperparameter selection. We introduce the concept of shrinkage factors for both the R2D2 and R2D2M2 prior as well as the relationship the prior's hyperparameters have with the effective number of nonzero coefficients in the model, thus extending the idea of Piironen and Vehtari (2017) to multilevel models. This provides an easy and intuitive way of setting up the prior, since it is possible to visualize how much sparsity will be induced. Section 4 shows the results of testing the R2D2M2 prior in intensive simulations. We make use of simulation based calibration (Talts et al., 2020) and provide evidence that that our implementation and estimation of both the R2D2 and the R2D2M2 prior models are well calibrated. What is more, we simulate data from sparse multilevel models and provide detailed discussions about how the prior behaves with respect to estimation error, out-of-sample predictive performance, credible interval coverage and Type I and Type II errors. Finally, we show how the R2D2M2 prior performs in real life data by testing it on the Riboflavin production data set made publicly available by Buehlmann et al. (2014). The paper ends with a discussion in Section 5 where we summarize our findings and suggest future lines of research. All mathematical proofs can be found in Appendix A.

2. The R2D2M2 prior

In the following, we denote the response variable as y, its nth observed value as y_n where $n \in \{1, ..., N\}$, and the predictor variables (features or covariates) as x_i with $i \in \{1, ...p\}$. Additionally we assume conditional independence of the N observations in the data. We define the overall coefficients as those that affect every member of the general population and the group or varying coefficients as those that do so per group. In the statistical literature, the terms overall and varying coefficients have been known as fixed and random effects, respectively, nevertheless we believe that this nomenclature is problematic in the Bayesian setting, since all of the parameters of interest are treated as (random) variables whose uncertainty is represented via probability distributions (Gelman and Hill, 2006). We denote the overall coefficients by $b_i, i \in \{0, ..., p\}$ and the group specific or varying coefficients by u_{ig_j}. Here g denotes a categorical grouping variable, encoding the individual groups g_j across which the coefficient of the ith predictor x_i is expected to vary. If we denote G_i as the index set of all grouping variables across which the coefficient of x_i is expected to vary, we can write the observation model of a linear multilevel model as

$$y_n \sim \text{Normal}\left(\mu_n, \sigma^2\right) \tag{1}$$

$$\mu_n = b_0 + \sum_{i=1}^{p} x_{ni} b_i + \sum_{g \in G_0} u_{0g_{j[n]}} + \sum_{i=1}^{p} x_{ni} \left(\sum_{g \in G_i} u_{ig_{j[n]}} \right), \tag{2}$$

for $n \in \{1, ..., N\}, i \in \{0, ..., p\}, g \in \{1, ..., G_i\}$ and $j \in J_g$, where J_g is the index set over the levels of grouping variable g. Here, b_0 and u_{0g_j} denote the overall intercept and varying intercepts, respectively, whereas the former will not be subject to extra prior regularization (see below).

When doing Bayesian inference in multilevel models, it is a common practice to set independent normal priors on all regression coefficients:

$$b_i \sim \text{Normal}\left(0, \tilde{\lambda}_i^2\right), \quad u_{ig_j} \sim \text{Normal}\left(0, \tilde{\lambda}_{ig}^2\right), \tag{3}$$

where, in current practice, $\tilde{\lambda}_i^2, \tilde{\lambda}_{ig}^2$ are either fixed or are assigned independent priors. The literature shows that either fixing the values of $\tilde{\lambda}$ or assigning independent priors for $\tilde{\lambda}$ is suboptimal, as the total prior variance of the model scales with the number of included terms, which is implausible in reality (Fuglstad et al., 2019). Beyond that, it remains unclear how these priors scale with the number of predictors when their coefficients vary across an increasing number of grouping factors.

2.1. Implied priors on R2

The R^2 measure expresses the proportion of variance explained by the model in relation to the total variance σ_y^2 of the response y (Gelman et al., 2020a). For

Gaussian models with residual standard deviation σ we define

$$R^2 \coloneqq \text{corr}^2(y, \mu) = 1 - \frac{\sigma^2}{\sigma_y^2}, \tag{4}$$

as a *global* measure of proportion of variance explained (Rights and Sterba, 2019). Global means that it jointly comprises all overall and group specific coefficients at the same time. This stands in contrast to some common definitions of R^2 in multilevel models, where R^2 is defined for each level of the model separately, rather than jointly across all levels and terms (Nakagawa and Schielzeth, 2013; Rights and Sterba, 2019). Our work emphasizes the use of a global R^2 as our aim is a joint regularization of all regression coefficients. In that sense, we use the above R^2 metric simply as a starting point for a joint regularizing prior for multilevel models, independently of which metric one would prefer to measure explained variances after estimating the model.

Given that R^2 is invariant under changes to the mean of the response y or any of the predictors x_i, we can assume $\mathbb{E}(y) = \mathbb{E}(x_i) = 0, \forall i \in \{1, ..., p\}$ without loss of generality in the following. This implies that the model's overall intercept b_{00} is zero and we can rewrite R^2 as

$$R^2 = \frac{\text{var}(\mu)}{\text{var}(\mu) + \sigma^2}. \tag{5}$$

Assume a-priori independence of the regression coefficients and consider a prior for b_i and u_{ig_j} that satisfies $\mathbb{E}(b_i) = 0, \mathbb{E}(u_{ig_j}) = 0$. Additionally, denote by $\sigma_{x_i}^2$ the variance of the ith covariate, then we can write

$$\text{var}(\mu) = \sum_{g \in G_0} \tilde{\lambda}_{0g}^2 + \sum_{i=1}^{p} \sigma_{x_i}^2 \left(\tilde{\lambda}_i^2 + \sum_{g \in G_i} \tilde{\lambda}_{ig}^2 \right). \tag{6}$$

Equation (6) shows that $\text{var}(\mu)$ can be decomposed into the variance components belonging to the overall and varying effects, respectively. Furthermore, the variance corresponding to the varying coefficients can be sub-decomposed into the variance associated to the varying intercepts and varying slopes.

The decomposition of $\text{var}(\mu)$ shown in Equation (6) allows us to investigate the effect the standard deviations $\tilde{\lambda}$ (or variances $\tilde{\lambda}^2$) have on the implied prior for R^2. As an example, consider a simple linear model without grouping variables $y_n \sim \text{Normal}(\mu_n, \sigma^2)$ where $\mu_n = \sum_{i=1}^{p} x_{ni} b_i$, with independent $\text{Normal}(0, \tilde{\lambda}_i^2)$ priors on b_i with $\tilde{\lambda}_i = 1$, and an Exponential(1) prior on σ. These type of priors are usually considered weakly informative when the variables are standardized, however Figure 1 shows that the implied prior on R^2 is actually very informative. The implied prior on R^2 becomes increasingly narrow with a peak close to $R^2 = 1$ as the number of predictors increases. This behavior becomes even stronger for larger values of λ_i and when adding more terms. Therefore, it is also present in the context of multilevel models, where including additional levels per grouping factor and/or additional grouping factors increases the amount of

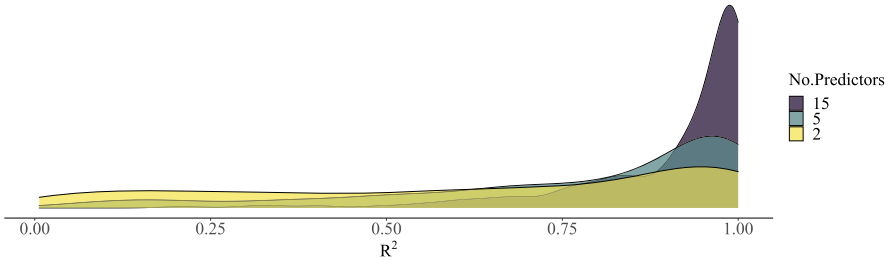


FIG 1. *Implied prior on R^2 varying number of predictors in a single level linear model with Normal$(0, 1)$ priors on the coefficients and an Exponential(1) prior on the error σ. Notice how weakly informative priors implicate a highly informative prior on R^2.*

parameters substantially. For instance, when considering one grouping factor g with L different levels and all p coefficients varying across the L levels, one would have $(p + 1)L$ additional coefficients (the 1 is for the varying intercept).

An $R^2 \approx 1$ represents almost perfect in-sample predictions. This is usually an indicator of overfitting when relevant amounts of noise are present in the data, such that $R^2 \approx 1$ can only be achieved by partially fitting on noise. Any prior implying such a large R^2 is highly unrealistic and would not be able to regularize sufficiently. Accordingly, it is more sensible to specify a prior on the overall variance explained by the model and then decompose this variance into components assigned to the individual terms, thereby inducing a joint prior over the variances of all terms. For this purpose, we propose (as have others (Goodrich et al., 2020; Zhang et al., 2020)) to specify a prior directly on R^2 as it provides an intuitive and widely understood measure of model fit.

As mentioned before, our main intention is to generalize the R2D2 prior proposed by Zhang et al. (2020) to models with increased complexity in terms of multilevel structures. On account of the inclusion of *M*ultilevel *M*odels (M2) in the R2D2 prior, we name our new prior the *R2D2M2 prior*.

2.2. Derivation of the R2D2M2 prior

In the following we present the derivation of our new prior. For this purpose, we begin by decomposing the priors' variances $\tilde{\lambda}_i^2$ and $\tilde{\lambda}_{ig}^2$ as

$$\tilde{\lambda}_i^2 = \frac{\sigma^2}{\sigma_{x_i}^2}\lambda_i^2, \quad \tilde{\lambda}_{ig}^2 = \frac{\sigma^2}{\sigma_{x_i}^2}\lambda_{ig}^2. \tag{7}$$

The factor $\frac{\sigma^2}{\sigma_{x_i}^2}$ accounts for the scales of y and x_i and ensures that λ_i represents prior variances on standardized variables, which are comparable across terms and λ_i are appropriate for any scale of y and x. In practice, σ_{x_i} is unknown and it could be replaced by the sample variance $\hat{\sigma}_{x_i}^2$ obtained from the data.

Plugging (7) into (6) results in

$$\text{var}(\mu) = \sigma^2 \left(\sum_{g \in G_0} \lambda_{0g}^2 + \sum_{i=1}^{p} \left(\lambda_i^2 + \sum_{g \in G_i} \lambda_{ig}^2 \right) \right) = \sigma^2 \tau^2,$$

where we have taken τ^2 as the sum of standardized prior variances given by

$$\tau^2 = \sum_{g \in G_0} \lambda_{0g}^2 + \sum_{i=1}^{p} \left(\lambda_i^2 + \sum_{g \in G_i} \lambda_{ig}^2 \right). \tag{8}$$

τ^2 is the (standardized) explained variance in correspondence to the interpretation of R^2 as proportion of explained variance. Using $\text{var}(\mu) = \sigma^2\tau^2$, R^2 can be written as

$$R^2 = \frac{\text{var}(\mu)}{\text{var}(\mu) + \sigma^2} = \frac{\sigma^2\tau^2}{\sigma^2\tau^2 + \sigma^2} = \frac{\tau^2}{\tau^2 + 1}. \tag{9}$$

Following Zhang et al. (2020), we set a Beta distribution on R^2 and we write $R^2 \sim \text{Beta}(\cdot, \cdot)$. The Beta distribution offers a wide number of different parametrizations (Kruschke, 2015). Here we choose to parametrize the Beta distribution on R^2 in terms of the prior mean μ_{R^2} and prior "precision" φ_{R^2} (also known as mean-sample size parameterization), since it provides users with an intuitive way to incorporate prior knowledge into the R^2 prior; Both hyperparameters are understandable and expressible in terms of domain knowledge that represent the existing relationship between the included covariates and the response variable.

Equation (9) implies that, by definition, τ^2 has a Beta-Prime distribution with parameters μ_{R^2}, φ_{R^2}, represented by $\tau^2 \sim \text{BetaPrime}(\mu_{R^2}, \varphi_{R^2})$. Figure 2 illustrates the flexibility that the Beta distribution can offer in expressing prior knowledge about R^2. The corresponding Beta-Prime prior for τ^2 is also shown. For example, for values of $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$, we get a bathtub-shaped prior on R^2 that places most mass near the two extremes $R^2 = 0$ and $R^2 = 1$. A priori, this indicates that the user expects the model to contain a substantial amount of either noise or signal.

In a consecutive decomposition step, we follow a Global-Local (GL) prior framework and set

$$\lambda_i^2 = \phi_i \tau^2, \quad \lambda_{ig}^2 = \phi_{ig} \tau^2, \tag{10}$$

with a vector ϕ containing all elements $\phi_i, \phi_{ig} \geq 0, \forall i \in \{1, ..., p\}, g \in G_i$ such that $\sum_i \phi_i + \sum_i \sum_{g \in G_i} \phi_{ig} = 1$, thus rendering ϕ as a simplex. ϕ_i and ϕ_{ig} represent the proportion of standardized explained variance attributed to the ith and igth term respectively. Decomposition (10) describes the proportion of standardized explained variance attributable to the igth term, therefore controlling the variability allocated to each individual variance λ_{ig}^2.

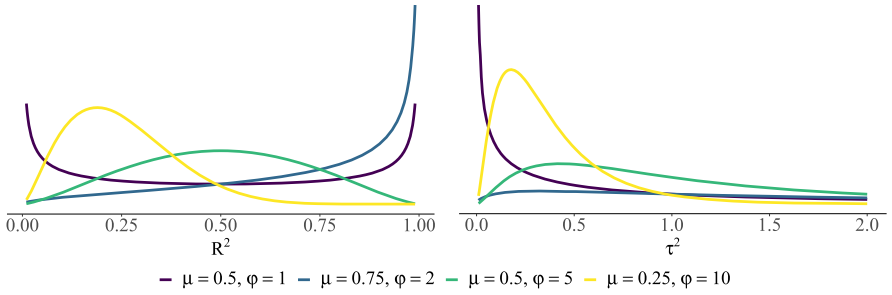


FIG 2. *Exemplary densities of the Beta prior for R^2 (left) and corresponding Beta-Prime prior for τ^2 (right) with varying mean and precision parameters μ, φ respectively.*

A natural prior for ϕ is a Dirichlet distribution with hyperparameter α chosen by the user, where α is a vector of length equal to the number of variance components that are included in the model and it governs the shape of the Dirichlet distribution. For instance, if the model has p overall coefficients, one grouping factor g and q varying coefficients in the grouping factor, then α is of length $p + q + 1$, due to the inclusion of the varying intercept. The amount of levels L does not affect the length of α according to our prior specification (e.g., see Figure 3 below).We denote the elements of α as α_i and α_{ig}, which represent the a priori expected relevance on each overall and varying regression term, respectively. In our context we will call α a concentration vector, since $\alpha_0 = \sum_i \alpha_i + \sum_{ig} \alpha_{ig}$ determines the narrowness of the distribution, that is, how peaked it is (Lin, 2016).

Naturally, the question arises of how α is to be specified. When $\alpha = (1, ..., 1)'$ the prior is flat over all possible simplexes, which renders this value a good choice in the absence of additional prior knowledge. More generally, setting $\alpha = (a_\pi, ..., a_\pi)'$ with a concentration parameter $a_\pi > 0$ (a symmetric Dirichlet distribution) can also be sensible since it drastically reduces the number of hyperparameters to specify, providing us with the ability to globally control the shape of the Dirichlet distribution with a single value. Additionally, the user can control the induced sparsity of the model by use of this single value as we show in Section 3.1. A symmetric Dirichlet distribution is useful when there is no prior preference that favours one component over another. Values for which $a_\pi < 1$ produce simplexes that concentrate their mass on the edges ("bathtub" distribution) and most coefficient values within a single sample will be close to 0. Instead, setting $a_\pi > 1$ results in concentration around the center of the simplex, favouring an even distribution among the different components (Lin, 2016). Finally, the user can also specify asymmetric Dirichlet distributions by designating different values for the components of α, which represent the differently a priori expected importance of the corresponding regression term.

The values of τ^2 and ϕ_{ig} play different roles in the prior. The explained variance τ^2 controls the amount of global shrinkage and is therefore deemed as

a global scale, whereas the igth attributed variance ϕ_{ig} serves as a local scale that controls the individual shrinkage the prior induces over each specific term. The components of ϕ compete to increase and to acquire a higher proportion from the total variance τ. This can be seen when examining the covariance between two different components, which is given by $\text{cov}(\phi_i, \phi_j) = \frac{-\alpha_i \alpha_j}{\alpha_0^2(\alpha_0+1)}, i \neq j$. Thus, for a fixed value of τ^2, as the importance of one term increases the importance of other terms decrease, and vice versa.

In the GL framework, the next step is to assign a base distribution that is able to express the attributed variance to each regression term b_i and u_{ig_j}. In their original R2D2 paper, Zhang et al. (2020) consider the coefficients to follow double exponential base distributions since they ensure both a higher concentration of mass around a zero and heavier tails than normal distributions. In the present paper, we will instead employ normal base distributions for each regression term for several, not mutually independent reasons:

1. It is customary in the multilevel model literature to consider normal base distributions for the overall and varying coefficients Gelman and Hill (2006); Wakefield (2013). What is more, Global-Local priors with normal base distributions are very flexible and have been the gold standard when proposing shrinkage priors, as can be seen in (Carvalho et al., 2010; Bhattacharya et al., 2015; Van Erp et al., 2019), among others.
2. Posterior approximation via gradient-based MCMC methods encounter issues when sampling from non-differentiable distributions, such as the double exponential (Brooks et al., 2011). Additionally, it is straighforward to reparametrize the posterior distribution to improve the rates of convergence of either Gibbs or Hamiltonian Monte Carlo based MCMC samplers in the context of conditionally normal multilevel models (Bürkner, 2017; Zanella and Roberts, 2021).
3. The use of normal distributions provides us with closed analytical expressions for several quantities of interest as described in Section 3 and shows, for example, good shrinkage and tail behavior.

To ease prior specification and speed up sampling, we will set a prior on the intercept $\tilde{b}_{00}$ implied when all $\mathbb{E}(x_i) = 0$ and then recover the original intercept b_{00} after model fitting using a simple linear transformation (Goodrich et al., 2020; Bürkner, 2017). Common priors on b_{00} are a normal prior with mean $\mathbb{E}(y)$ and a user chosen scale depending on the scale of y as well as Jeffrey's prior which is improper flat in this case (Good, 1962).

What remains to be specified is a prior for the residual variance σ^2 (or equivalently residual standard deviation σ). We follow the recommendations of Gelman (2006) to consider a half Student-t prior on σ with ν degrees of freedom and scale η. We propose setting $\eta \approx \text{sd}(y)$, since both the prior expected mean and variance are proportional to η. To be able to implement a Gibbs sampling approach, an inverse Gamma distribution $\sigma^2 \sim \text{InvGamma}(c, d)$, with shape and scale parameters c, d respectively, could also be considered. We show how to implement a Gibbs sampler for the R2D2M2 prior in Appendix B.

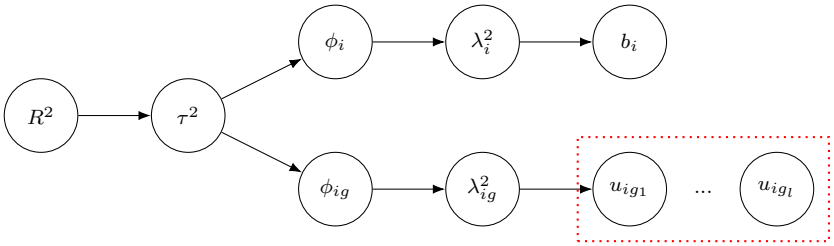


FIG 3. *Construction of the R2D2M2 prior schematically. We begin by assigning a distribution to R^2, afterwards its uncertainty is propagated to the regression terms via a Dirichlet decomposition of the (a priori) explained variance. Finally, we make use of a distribution that can express the amount of variance that has been allocated to each term. Notice how all u_{ig_jl} share the same local variance λ_{i_g} for a given pair (i, g_j) and varying level l.*

The full R2D2M2 model can summarized as

$$
\begin{aligned}
y_n &\sim \text{Normal}(\mu_n, \sigma^2) \\
\mu_n &= b_0 + \sum_{i=1}^{p} x_{ni} b_i + \sum_{g \in G_0} u_{0g_{j[n]}} + \sum_{i=1}^{p} x_{ni} \left(\sum_{g \in G_i} u_{ig_{j[n]}} \right) \\
b_0 &\sim p(b_0) \\
b_i &\sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2} \phi_i \tau^2\right), \quad u_{ig_j} \sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2} \phi_{ig} \tau^2\right) \\
\tau^2 &= \frac{R^2}{1 - R^2} \\
R^2 &\sim \text{Beta}(\mu_{R^2}, \varphi_{R^2}), \quad \phi \sim \text{Dirichlet}(\alpha), \quad \sigma \sim p(\sigma).
\end{aligned}
\tag{11}
$$

A schematic construction of the R2D2M2 is shown in Figure 3. We do not specify distributions in this diagram, thus emphasizing that there is flexibility in the way uncertainty can be propagated from R^2 to the coefficients (as long as consistent rules that respect the corresponding domains of each parameter are followed). For example, the user might want to use a different distribution for R^2, such as the Kumaraswamy distribution which offers advantages when using quantile-based approach to statistical modelling (Jones, 2009), or they might desire to model correlations among the components of ϕ explicitly, in which case they could use a distribution that allows it, such as a logistic normal distribution (Aitchison and Shen, 1980).

We close this section commenting on how the intuitiveness of the R2D2M2 prior assists in eliciting prior knowledge or lack of it. Users can readily communicate a prior expected value for R^2 through μ_{R^2}. Given a value for μ_{R^2}, the value selected for φ_{R^2} is an indicator of how confident the user is on how much of the response variability the model is explaining. This falls in line with the recommendations given by Mikkola et al. (2021) on how to improve prior elicitation: the hyperparameters on the prior for R^2 are easy to understand and

the priors for the regression terms regularized jointly (see Figure 3), hence significantly decreasing the number of hyperparameters to the user has to choose. Thus, the R2D2M2 prior brings us a step forward in terms of prior elicitation for multilevel models and provides a valuable tool in the Bayesian workflow for data analysis more generally (Gelman et al., 2020b).

3. Properties of the R2D2M2 prior

In this section, we derive several mathematical properties of the R2D2M2 prior. We begin by presenting the marginal prior distributions of the overall and varying coefficients. We show how the behavior around the origin and the tails of the marginal densities can be controlled by the hyperparameters $\mu_{R^2}, \varphi_{R^2}, a_\pi$. The distinct behaviors induced by hyperparameter choices can be exploited by the user to represent their prior knowledge and desired degree of regularization.

We then present the conditional posterior distributions of the overall and varying coefficients. Specifically, based on their conditional posterior means, we introduce shrinkage factors that allow us to calculate the effective number of coefficients and give further guidelines on how to choose the hyperparameters of the R2D2M2 prior.

3.1. Tail and concentration behaviors

The original R2D2 prior proposed by Zhang et al. (2020) uses a double exponential (Laplace) base distribution for the coefficients b_i. However, as mentioned before, in the multilevel model literature it is customary to treat b_i and in particular the varying coefficients u_{ig_j} as normally distributed (see Section 2 for details). The use of normal base distributions for the coefficients may lead to different marginal prior distributions, different concentration properties around the origin and tail behavior, which we will study in the following.

Consider the hyperparameters of the R2D2 prior μ_{R^2}, φ_{R^2} and let $a_1 = \mu_{R^2}\varphi_{R^2}, a_2 = (1-\mu_{R^2})\varphi_{R^2}$ be the corresponding shape parameters of the Beta distribution on R^2 in the standard parameterization. In the following choose $\alpha = (a_\pi, ..., a_\pi)'$ with $a_\pi > 0$. Zhang et al. (2020) showed that, under specific conditions depending on the hyperparameters a_π and a_2, their prior is able to attain polynomial concentration rates around the origin and polynomial decay rates in the tails, thus allowing for shrinkage of noise towards zero and detection of signals. This shows that the R2D2 prior is able to compete with other well known GL shrinkage priors. We arrive to similar conclusions when using normal base distributions and taking multilevel structure into consideration. In the following propositions, the marginal distributions presented depend on a fixed value of σ. Other authors (Carvalho et al., 2010; Bai and Ghosh, 2019; Zhang et al., 2020) usually fix the value of $\sigma = 1$ for simplicity, however the marginal priors they present are still conditional on σ, as are ours.

Proposition 3.1. *Given the R2D2M2 prior, the marginal prior densities of* b_i, u_{ig_j} *given* σ *for any* $i = 1, ..., p, g \in \{1, ..., G_i\}, j \in J_g$ *are*

$$p(b_i|\sigma) = \frac{1}{\sqrt{2\pi q_i^2} B(a_\pi, a_2)} \Gamma(\eta) U(\eta, \nu, z_i) \tag{12a}$$

$$p(u_{ig_j}|\sigma) = \frac{1}{\sqrt{2\pi q_i^2} B(a_\pi, a_2)} \Gamma(\eta) U(\eta, \nu, z_{ig_j}), \tag{12b}$$

where $q_i^2 = \frac{\sigma^2}{\sigma_{x_i}^2}, a_2 = (1 - \mu_{R^2})\varphi_{R^2}, \eta = a_2 + 1/2$ *and* $\nu = 3/2 - a_\pi$. $B(\cdot, \cdot)$ *and* $\Gamma(\cdot)$ *represent the Beta and Gamma functions respectively and* $U(\eta, \nu, z_i)$ *represents the confluent hypergeometric function of the second kind, as well as* $z_i = \frac{|b_i|^2}{2q_i^2}$ *and* $z_{ig_j} = \frac{|u_{ig_j}|^2}{2q_i^2}$. *If there are no varying coefficients, the marginal prior densities of* b_i *remain unaltered.*

Proposition 3.2. *As* $|b_i| \to 0, |u_{ig_j}| \to 0$, $0 < a_\pi \leq 1/2$, *and* $a_2 > 0$, *the marginal prior densities are unbounded with a singularity (i.e., undefined) at zero. Moreover, the marginal priors satisfy*

$$p(b_i|\sigma) \sim \begin{cases} c_1 b_i^{2a_\pi - 1} + \mathcal{O}\left(|b_i|^{2a_\pi + 1}\right), & a_\pi < 1/2 \\ -c_2 \ln(b_i^2/2q_i^2) + \mathcal{O}\left(b_i^2 \ln(b_i^2/2q_i^2)\right), & a_\pi = 1/2 \end{cases}$$

$$p(u_{ig_j}|\sigma) \sim \begin{cases} c_1 u_{ig_j}^{2a_\pi - 1} + \mathcal{O}\left(|u_{ig_j}|^{2a_\pi + 1}\right), & a_\pi < 1/2 \\ -c_2 \ln(u_{ig_j}^2/2q_i^2) + \mathcal{O}\left(u_{ig_j}^2 \ln(u_{ig_j}^2/2q_i^2)\right), & a_\pi = 1/2 \end{cases}$$

where $c_1 = (2q_i^2)^{1/2 - a_\pi} \Gamma(1/2 - a_\pi)/\Gamma(a_2 + 1/2), c_2 = 1/\Gamma(a_2 + 1/2)$. *When instead* $a_\pi > 1/2$, *the marginal prior densities are bounded and continuous at zero. The marginal prior densities are differentiable at* $t = 0$ *for all values of* $a_\pi \geq 1$.

Proposition 3.2 implies that the value of a_π controls the boundedness of the marginal priors near the origin, which in turn influences the level of shrinkage that is enforced upon the coefficients. Sufficient mass near zero allows the prior to shrink small signals sufficiently (Van Erp et al., 2019). As a_π decreases from 1/2 to zero, the R2D2M2 prior shifts its mass even stronger to the origin and we can expect the posterior distribution to be concentrated near zero. This leads to prior concentration around sparse coefficient vectors and favors sparse estimation of the coefficients. On the other hand, an increasing a_π from 1/2 produces a bounded prior and indicates that we are in favor of less sparse coefficient vectors.

We can further understand the behavior of the marginal priors by considering that, when assuming $\phi \sim \text{Dirichlet}(\alpha)$ where $\alpha = (a_\pi, ..., a_\pi)'$, the value of a_π determines where the density places its mass (Lin, 2016). When $0 < a_\pi < 1$ the density congregates at the edges of the simplex and prefers sparse distributions, i.e., most of the values within a single sample will be close to 0, and the vast majority of the mass will be concentrated in a few of the ϕ values. As a_π increases

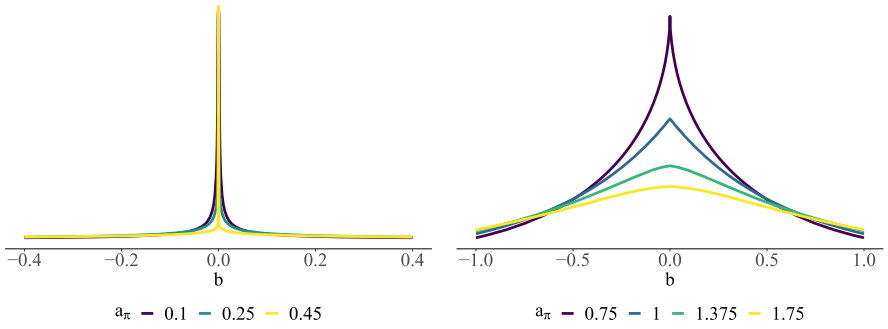


FIG 4. *Marginal prior densities of the coefficients with* $\sigma = 1$. *The value of* a_π *controls the behavior around the origin. If* $a_\pi \leq 1/2$ *the marginal density increases towards infinity for values close to zero and is unbounded and non-differentiable at the origin. In contrast, if* $a_\pi > 1/2$ *the marginal density is bounded also at the origin. When* $a_\pi \geq 1$ *the densities become differentiable at the origin.*

to and exceeds 1, the density concentrates near the center of the simplex, with a mode appearing in $\left(\frac{1}{\dim(\alpha)}, ..., \frac{1}{\dim(\alpha)}\right)'$, where $\dim(\alpha)$ denotes the length of α. Values of $a_\pi > 1$ prefer variates that are dense and evenly distributed, i.e., all the values within a single sample are similar to each other. This property is inherited by the R2D2M2 prior as Proposition 3.2 shows, since selecting $a_\pi \leq 1/2$ will attempt to concentrate the total variance into few regression terms, whereas $a_\pi > 1/2$ will distribute it evenly among the terms. Translation of prior beliefs over sparsity or lack thereof is thus possible by selection of the parameter a_π.

Figure 4 shows the marginal density of $p(b_i|\sigma)$ for different values of a_π with fixed $q_i^2 = 1$, the latter implying an equal scale of predictors and response variable without loss of generality. For the value of $a_\pi = 0.25$, the prior distribution shows high concentration of mass near the origin and is undefined (singular) at zero, which will lead to stronger shrinkage of weak signals towards the origin. On the other hand, when $a_\pi = 0.75$, the singularity is not present anymore and the prior becomes bounded at zero. A case of interest that presents bounded marginal priors is that of the (uniform) flat Dirichlet distribution over the simplex, that is when $a_\pi = 1$, meaning that the use of the flat Dirichlet prior over ϕ enforces less sparsity. Both types of behavior can be beneficial, since they can play in favor of the prior beliefs about sparsity the user might have.

The tails of the prior dictate whether or not it is possible that large signals can be detected (even when specific levels of sparsity are assumed) and how much these are shrunken towards zero (Carvalho et al., 2010; Bhattacharya et al., 2015; van der Pas, 2021). It is in principle desirable that heavy tails are present in order to prevent over-shrinkage of true signals (Carvalho et al., 2010; Piironen and Vehtari, 2017), which in turn can result in priors that are of bounded influence, i.e., sufficiently strong signals are left unshrunken by the prior. This behavior, known as tail robustness (Carvalho et al., 2010; van der Pas, 2021), is vital in sparse settings to be able to shrink coefficients near zero

much more forcefully than those far from it. However, we should keep in mind that in practice, this behavior does not always plays in our favour, especially when parameters are weakly identified by the data. To circumvent this issue, authors such as Piironen and Vehtari (2017); Nishimura and Suchard (2022) have proposed *regularized* versions of shrinkage priors that also regularize very strong signals to some degree.

Our proposed R2D2M2 prior is able to attain heavier tails than the Cauchy distribution and, when no varying terms are present (the R2D2 prior with normal base distributions), it is of bounded influence, as is shown in the following propositions.

Proposition 3.3. *Given* $|b_i| \to \infty, |u_{ig_j}| \to \infty$ *for any* $a_\pi > 0$ *and* $a_2 > 0$ *the marginal prior densities satisfy*

$$p(b_i|\sigma) \sim \mathcal{O}(1/|b_i|^{2a_2+1}), \quad p(u_{ig_j}) \sim \mathcal{O}(1/|u_{ig_j}|^{2a_2+1}).$$

Moreover, when $0 < a_2 \leq 1/2$ *the R2D2M2 marginal priors have heavier tails than the Cauchy distribution.*

Proposition 3.4. *Assume there are no varying coefficients* u_{ig_j} *in the model, then the R2D2M2 prior is of bounded influence, i.e., sufficiently strong signals are not shrunken.*

While we were not able to prove that bounded influence also holds for the R2D2M2 prior, we have no reason to believe that it does not hold. Indeed, the results from the simulations presented in Section 4 show the bounded influence for the R2D2M2 prior empirically, at least for a range of different hyperparameter and data generating setups.

Propositions 3.2 and 3.3 show the adaptability of the R2D2M2 prior, in the sense that the user can tune the hyperparameters according to their prior beliefs on the degree of sparsity and signal strength. However, one should keep in mind that the complexity of the problem rapidly increases when considering multilevel structures and care should be taken when selecting the hyperparameters as to neither overfit (too little sparsity) nor over-shrink (too much sparsity).

3.2. Shrinkage factors

When performing Bayesian inference, the prior can be seen as a regularizer. This is clearly exemplified when using the Bayesian LASSO or other Bayesian shrinkage priors (Park and Casella, 2008; Van Erp et al., 2019), since the posterior modes under these priors correspond to certain frequentist regularizers (Bhattacharya et al., 2015). However, usual frequentist properties are not precisely equivalent; for instance, the posterior distribution for a continuous random variable is not able to produce exact sparse results, as the latter would require a point mass at zero. When using shrinkage priors, some authors have been able to express the posterior means as shrunken versions of (unshrunken) frequentist estimators (Piironen and Vehtari, 2017), which in practice leads to a better

out-of-sample predictive performance and a decrease in the variance of the estimators (Hoerl and Kennard, 1970). Several authors have used the concept of *shrinkage factors* to study the amount of shrinkage the prior induces and the effect the different values of the hyperparameters (Carvalho et al., 2010; Piironen and Vehtari, 2017; Bai and Ghosh, 2019; Polson and Scott, 2012). Here, we present the theory of shrinkage factors for the R2D2M2 prior as well as for the original R2D2 prior, since it has not been done before.

It can be shown that the conditional posterior distributions of the coefficients b_i and u_{ig_j} given $b_{-i}, u_{-ig_j}, \phi, \sigma, \tau, y$ in (11) are normal with means given by

$$\begin{aligned}\mathbb{E}(b_i|b_{-i}, u, \sigma, \phi, \tau^2, y) &= \left(\frac{\sigma^2_{x_i}}{\phi_i\tau^2} + SS_i\right)^{-1} \sum_{n=1}^{N} e_n^{(-i)} x_{ni} \\ \mathbb{E}(u_{ig_j}|b, u_{-ig_j}, \sigma, \phi, \tau^2, y) &= \left(\frac{\sigma^2_{x_i}}{\phi_{ig}\tau^2} + SS_{ig_j}\right)^{-1} \sum_{n\in L_{g_j}} e_n^{(-ig_j)} x_{ni},\end{aligned} \tag{13}$$

where b_{-i} and u_{-ig_j} represent the exclusion of the ith overall coefficient and the ig_j varying coefficient, respectively, $SS_i = \sum_n x_{ni}^2$ and $SS_{ig_j} = \sum_{n\in L_{g_j}} x_{ni}^2$. We use $e_n^{(-i)}$ and $e_n^{(-ig_j)}$ to represent the errors that would be obtained when excluding b_i and u_{ig_j} when trying to predict the nth observation.

$$\begin{aligned}e_n^{(-i)} &= y_n - \sum_{i'\neq i} x_{ni'} b_{i'} - \sum_{i=0}^{p} x_{ni} \left(\sum_{g\in G_i} u_{ig_{j[n]}}\right) \\ e_n^{(-ig_j)} &= y_n - \sum_{i=1}^{p} x_{ni} b_i - \sum_{i'\neq i} \sum_{g\in G_{i'}} x_{ni'} u_{i'g_{j[n]}} - x_{ni} \sum_{g\in G_i, g'\neq g} u_{ig'_{j[n]}}.\end{aligned}$$

If we assume there are no varying terms in the model, i.e., we are considering the original R2D2 prior, and $\sum_n x_{ni}x_{nj} = 0$ for $i \neq j$ then the conditional posterior mean of b_i is given by

$$\mathbb{E}(b_i|\sigma, \phi, \tau^2, y) = \left(\frac{\sigma^2_{x_i}}{\phi_i\tau^2} + SS_i\right)^{-1} SS_i\, \hat{b}_i, \tag{14}$$

where $\hat{b}_i$ is the maximum likelihood estimate (MLE) of b_i. Thus, the posterior mean under the R2D2 prior is a shrunken version of the MLE, a common behavior that is encountered when regularizing (Hoerl and Kennard, 1970) or performing Bayesian inference with shrinkage priors such as the regularized horseshoe (Piironen and Vehtari, 2017). We are able to quantify the amount of shrinkage of b_i from their MLE towards zero by introducing shrinkage factors denoted as κ_i, where $0 \leq \kappa_i \leq 1$. Let

$$\kappa_i = \frac{1}{1 + \frac{SS_i}{\sigma^2_{x_i}}\phi_i\tau^2} \tag{15}$$

then Equation (14) becomes

$$\mathbb{E}(b_i|\sigma, \kappa_i, y) = (1 - \kappa_i)\hat{b}_i. \tag{16}$$

The shrinkage factor κ_i quantifies the amount of shrinkage that is exerted by the use of the R2D2 prior, relative to the non-regularized MLE. Shrinkage values close to 1 indicate full shrinkage from the MLE towards 0 and vice-versa, values close to 0 indicate no or only minor shrinkage.

When considering the presence of varying terms in the model, frequentist inference for b_i is typically performed by first integrating out u_{ig_j} (Bates et al., 2015). From a practical perspective, this is necessary as it is not possible to fully identify the overall and varying coefficients simultaneously when solving the optimization problem related to maximum likelihood. Thus, classical inference chooses to predict, rather than estimate, the varying terms u_{ig_j} after having obtained estimates $\hat{b}_i$ of the overall coefficients b_i (Bates et al., 2015; Wakefield, 2013). In Bayesian inference on the other hand, both unknown quantities b_i and u_{ig_j} are considered random variables and are present in the full posterior distribution. Considering this, the generalization of shrinkage factors from an MLE when including varying terms is not immediate, because the joint MLE of the b_i and u_{ig_j} does not exist in this case. However, Equations (13) still show shrinkage is being carried out from some fixed quantities that involve the conditioned parameters. The approach to define shrinkage from quantities different than the MLE has been considered by other authors too (Bai and Ghosh, 2019; Polson and Scott, 2012), although not in a multilevel context. Doing so here opens up the way to define shrinkage factors κ_i and κ_{ig_j} for overall and varying coefficients respectively by

$$\kappa_i = \frac{1}{1 + r_i\phi_i\tau^2}, \qquad \kappa_{ig_j} = \frac{1}{1 + r_{igj}\phi_{ig}\tau^2}, \tag{17}$$

where $r_i = \frac{SS_i}{\sigma^2_{x_i}}$ and $r_{ig_j} = \frac{SS_{ig_j}}{\sigma^2_{x_i}}$. Substituting in Equations (13) results in

$$\begin{aligned} \mathbb{E}(b_i|b_{-i}, \sigma, \kappa_i, y) &= (1 - \kappa_i)(SS_i)^{-1} \sum_{n=1}^{N} e_n^{(-i)} x_{ni} \\ \mathbb{E}(u_{ig_j}|b, u_{-ig_j}, \sigma, \kappa_{igj}, y) &= (1 - \kappa_{ig_j})(SS_{ig_j})^{-1} \sum_{n \in L_{g_j}} e_n^{(-ig_j)} x_{ni}. \end{aligned} \tag{18}$$

Shrinkage factors are useful in providing insight on the influence of the prior hyperparameters μ_{R^2}, φ_{R^2}, α, and a_π on the posterior distribution of each term. For instance, notice that in the extreme cases that α is chosen such that for a fixed i, $\phi_i \to 0$, then $\kappa_i \to 1$. If $\phi_i \to 1$ we have $\kappa_i \to (1 + r_i\tau^2)^{-1}$, so regularization takes place even if a priori it is believed that all the explainable variance is explained by a single component b_i. Additionally, by comparing κ_i and κ_{igj} in Equation (17) we can see that if $\phi_i \approx \phi_{ig}$ (which can be expected a-priori if $\alpha = a_\pi(1, ..., 1)'$ when $a_\pi > 1$ (Lin, 2016)), then the varying coefficients will exhibit stronger shrinkage since $SS_{ig_j} \leq SS_i$ naturally, due to observations being

partitioned into different levels. It is also worth mentioning that the shrinkage phenomenon present in the posterior of u_{ig_j} is two-fold: (1) resulting from the usual shrinkage of varying coefficients that hierarchical models exhibit (i.e., partial pooling, Gelman and Hill (2006)) and (2) resulting from the local and global scales that are part of the R2D2M2 prior.

It is possible to find a closed expression for the densities of κ_i and κ_{ig_j}, as well as for their moments. Consider Equations (17) with ϕ fixed and τ^2 random. Using the fact that $\tau^2 \sim \text{BetaPrime}(\mu_{R^2}, \varphi_{R^2})$, the prior densities of the shrinkage factors, conditional on ϕ are given by

$$\begin{aligned} p\left(\kappa_i|\phi_i\right) &= \frac{(r_i\phi_i)^{a_2}}{\mathrm{B}(a_1,a_2)}(1-\kappa_i)^{a_1-1}\kappa_i^{a_2-1}\left((r_i\phi_i-1)\kappa_i+1\right)^{-a_1-a_2}, \\ p\left(\kappa_{ig_j}|\phi_{ig}\right) &= \frac{(r_{ig_j}\phi_{ig})^{a_2}}{\mathrm{B}(a_1,a_2)}(1-\kappa_{ig_j})^{a_1-1}\kappa_{ig_j}^{a_2-1}\left((r_{ig_j}\phi_{ig}-1)\kappa_{ig_j}+1\right)^{-a_1-a_2}. \end{aligned} \tag{19}$$

The m-th moments of $\kappa_i|\phi_i$ and $\kappa_{ig_j}|\phi_{ig}$ with $m \in \mathbb{Z}^+$ are given respectively by

$$\begin{aligned} \mathbb{E}\left(\kappa_i^m|\phi_i\right) &= \frac{(r_i\phi_i)^{a_2}}{\mathrm{B}(a_1,a_2)}\mathrm{B}(a_2+m,a_1)\,{}_2F_1\left(\xi,\beta,\gamma,z_i\right), \\ \mathbb{E}\left(\kappa_i^m|\phi_i\right) &= \frac{(r_{ig_j}\phi_{ig})^{a_2}}{\mathrm{B}(a_1,a_2)}\mathrm{B}(a_2+m,a_1)\,{}_2F_1\left(\xi,\beta,\gamma,z_{ig_j}\right), \end{aligned}$$

where ${}_2F_1\left(\xi,\beta,\gamma,z\right)$ is the hypergeometric function (Jeffrey et al., 2007) with $\xi = a_1+a_2, \beta = a_2+m, \gamma = a_1+a_2+m$, $z_i = 1-r_i\phi_i$ and $z_{ig_j} = 1-r_{ig_j}\phi_{ig}$. In the important case of $m=1$, we have

$$\begin{aligned} \mathbb{E}\left(\kappa_i|\phi_i\right) &= (r_i\phi_i)^{a_2}\left(1-\mu_{R^2}\right){}_2F_1\left(\xi,\beta,\gamma,z_i\right), \\ \mathbb{E}\left(\kappa_{ig_j}|\phi_{ig}\right) &= (r_{ig_j}\phi_{ig})^{a_2}\left(1-\mu_{R^2}\right){}_2F_1\left(\xi,\beta,\gamma,z_{ig_j}\right). \end{aligned} \tag{20}$$

The distribution of the shrinkage factors κ depends directly on the hyperparameters μ_{R^2}, φ_{R^2} specified for the prior on R^2 and is sufficiently flexible to attain different forms that represent a diversity of a priori beliefs as we show in Figure 5. For example, the prior density of κ can become unbounded near 0 or 1 (or both at the same time), favoring limited or full shrinkage respectively. It is also possible to obtain bounded densities, which can represent mild shrinkage. These differences influence how noise and large signals are treated, which is reflected in the values of posterior means (Carvalho et al., 2010). Equations (20) show the effect of μ_{R^2} on the amount of shrinkage and how the prior is able to relate prior belief on R^2 with expected shrinkage on the coefficients. The expected shrinkage of each individual coefficient in the model is proportional to $1-\mu_{R^2}$, implying that μ_{R^2} acts in a global manner. This is expected due to the relationship between the model's global scale τ^2 and R^2 established in Equation (9). Prior beliefs about R^2 are straightforwardly inherited to the amount of shrinkage the prior exhibits. If the user considers R^2 is low and represents this via a low value of μ_{R^2}, then a noticeable amount of shrinkage will take place. Analogously, a high value of μ_{R^2} results in less shrinkage.

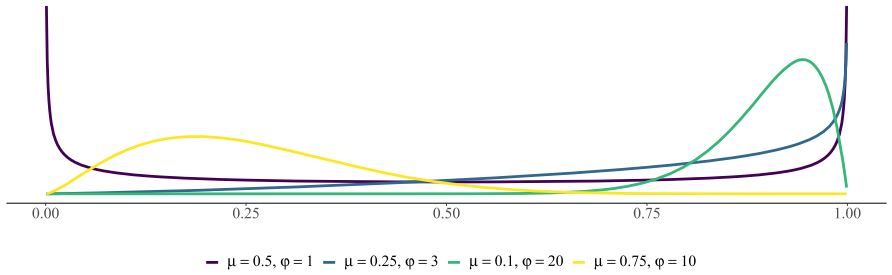


FIG 5. *Densities of shrinkage factors κ for several hyperparameter choices. The distribution offers flexibility to express different shrinkage behaviors. Notice that when $\mu = 0.5$ and $\varphi = 1$ the distribution is symmetric, categorizing the coefficient as either complete noise ($k \approx 1$) or as a strong signal ($k \approx 0$).*

Figure 5 shows that the density of κ is able to reflect a horseshoe like prior for the amount of shrinkage when $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$. This type of prior can be considered fairly non-informative on the κ scale, since it places $1/3$ of its mass on $1/4 \leq \kappa \leq 3/4$, with the rest of the density equally distributed to the left of $1/4$ and to the right of $3/4$. This shape indicates that we expect to see both relevant strong signals ($\kappa \approx 0$) and irrelevant variables a priori ($\kappa \approx 1$) (Carvalho et al., 2010; Piironen and Vehtari, 2017).

3.3. *Effective number of non-zero coefficients*

Piironen and Vehtari (2017) proposed that the prior's hyperparameters can be understood intuitively by analyzing the imposed prior on the *effective number of coefficients*. In the case of single level models they define the effective number of overall coefficients as

$$m_{\text{eff}_O} = \sum_{i=1}^{p} (1 - \kappa_i). \tag{21}$$

We generalize Equation (21) to multilevel models (11) by also including the varying coefficients:

$$m_{\text{eff}} = \sum_{i=1}^{p} (1 - \kappa_i) + \sum_{i \in \{0,\ldots,p\}} \sum_{g \in G_i} \sum_{j \in J_g} (1 - \kappa_{ig_j}). \tag{22}$$

When the shrinkage factors κ are close to 0 or 1, resulting in no shrinkage and total shrinkage respectively, Equation (22) can help us understand how specific values for μ_{R^2}, φ_{R^2} determine the amount of unshrunken ("active") variables present in the model and can serve as an indicator of the effective model complexity.

For a given fixed dataset, we can visualize the prior imposed on m_{eff} for different hyperparameter values of μ_{R^2}, φ_{R^2}. This can be done by simulating

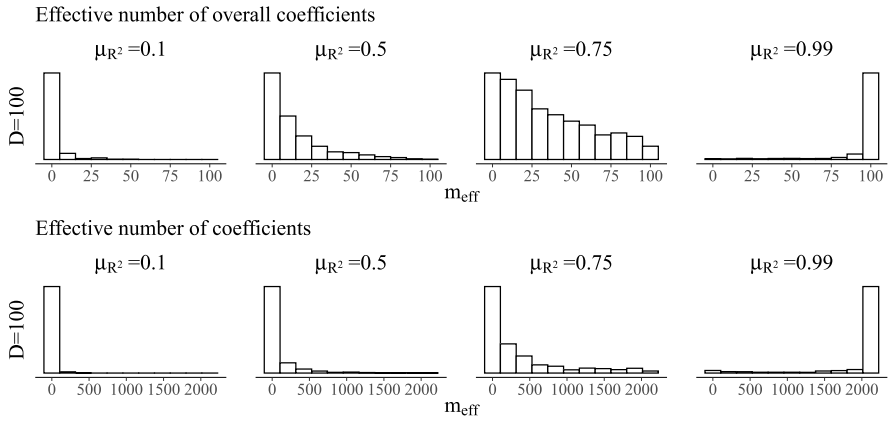


FIG 6. *Prior densities for the effective number of coefficients under different conditions for* μ_{R^2} *with fixed* $\varphi_{R^2} = 1, a_\pi = 0.5$ *and* $p = 100$ *overall coefficients, one grouping factor and 20 levels. Decreasing* μ_{R^2} *implies less signal expected a priori, which results in increased a priori shrinkage of the coefficients. These plots serve as an intuitive way to understand the effect the hyperparameters have on the amount of shrinkage.*

observations from m_{eff} directly. To do so, we first generate random variates τ, ϕ independently from their prior distributions, then we compute κ_i, κ_{ig_j} from Equations (17) and finally calculate m_{eff}. Figure 6 shows the histograms of the effective number of overall coefficients and total effective number of coefficients (which include the overall plus varying coefficients). In both cases, the behavior is as previously described; as μ_{R^2} increases the effective number of overall coefficients increases too and vice versa.

We recommend that the user visualizes the prior of m_{eff} for different values of μ_{R^2}, φ_{R^2} to get a better idea of how the global shrinkage is affected by the selection of the hyperparameters; and to verify if this choice represents their prior expectations on the number of non-zero coefficients. The prior plots of m_{eff} provide an intuitive way of understanding the induced overall shrinkage and can even become a valuable tool of communication between the user and non-specialized audiences.

4. Numerical experiments and case studies

We have implemented the R2D2M2 model in the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2022). The corresponding code can be found in Appendix B as well as on `https://osf.io/wgsth/`. We have also implemented the model in the R package `brms` (Bürkner, 2017), which provides a high level interface to fit Bayesian (non-)linear multivariate multilevel models using Stan. Based on the pure Stan implementation, we performed two large-scale simulations studies and analysed a high-dimensional real-world data set using the R2D2M2 prior as detailed below.

4.1. Simulation based calibration

First, we demonstrate that inference for our implementation of the R2D2M2 prior is probabilistically well calibrated. For this purpose, we applied simulation based calibration (SBC) (Talts et al., 2020). SBC is a general procedure for validating inferences from Bayesian algorithms capable of generating samples from posterior distributions. The procedure gives us the ability to identify inaccurate computations as well as inconsistencies in model implementations. In works by first sampling true parameter values $\tilde{\theta}$ from the prior, $\tilde{\theta} \sim p(\theta)$ and subsequently sampling data $\tilde{y}$ from the likelihood $\tilde{y} \sim p(y|\tilde{\theta})$. Afterwards, using the algorithm that needs to validated, we obtain S samples $\{\theta_1, ..., \theta_S\} \sim p(\theta|\tilde{y})$ from the (approximated) posterior. Finally, for a quantity of interest $f(\theta)$, we calculate a rank statistic of how many values in $\{f(\theta_1), ..., f(\theta_n)\}$ fall below the true value $f(\tilde{\theta})$. If the algorithm is well calibrated, this rank statistic of the prior sample $f(\tilde{\theta})$ in relation to the posterior draws is uniformly distributed (Talts et al., 2020). We will use the terms posterior draws and samples interchangeably.

To test uniformity of the rank statistics, we applied a intuitive graphical test proposed by Säilynoja et al. (2021), which provides simultaneous confidence bands of the empirical cumulative distribution function (ECDF) that satisfy a pre-specified joint coverage assuming uniformity. For ease of illustration, the plots show differences between the perfectly uniform CDF (a diagonal line) and the ECDF. In other words, it rotates a regular ECDF plot by 45 degrees to the right to make the uniform CDF a flat line and make the relevant part of the plot stand out more clearly. The graphical test is straightforward to use and is implemented in the R package `SBC` (Kim et al., 2022).

To sample from the posterior distribution we use Stan (Carpenter et al., 2017; Stan Development Team, 2022), a probabilistic programming language that provides the user with a (now substantially extended) implementation of the No-U-Turn Sampler (NUTS) from Hoffman and Gelman (2014), an adaptive form of Hamiltonian Monte Carlo (HMC) sampling (Brooks et al., 2011). To avoid sampling issues often encountered in hierarchical models we use non-centered parametrizations of the model (see `https://osf.io/wgsth/` and Gorinova et al. (2019) for details).

Calibration is a function of both the probabilistic model under consideration and the posterior approximation algorithm. Here we have tested calibration of our Stan implementations of both the R2D2 and the R2D2M2 prior models, since both of them are highly practically relevant. Note that calibration tests for previous implementations of the R2D2 prior model have not been reported beforehand by other authors.

There are two quantities that play a key role in how computationally intensive SBC can turn out to be: the number of simulation trials per configuration (i.e., the number of fitted models per configuration) and how many posterior draws to extract per simulation (i.e., per fitted model). We have fixed these to $T = 100$ and $S = 3000$, respectively. For MCMC samplers such as random walk Metropolis-Hastings or Gibbs sampling, this amount of posterior draws might seem as a bit small, however for adaptive Hamiltonian Monte Carlo estimates,

a few thousand samples is sufficient for most models (Bürkner, 2017; Hoffman and Gelman, 2014; Stan Development Team, 2022); and, as we show below, is indeed sufficient for the here-considered models.

We have tested calibration for a total of 96 different simulation configurations as detailed below. This amount comes from fully crossing all the factors that are varied in the simulations. An overview of the different configurations can be found in Appendix C in Table 5. The different conditions are chosen to represent several realistic scenarios that analysts might encounter in a similar fashion in real-life data analysis.

To sample prior values $\tilde{\theta}$ from the R2D2 or the R2D2M2 prior we need to specify several quantities: the prior mean μ_{R^2} and prior precision φ_{R^2} of the prior on R^2 as well as the concentration parameter a_π of the Dirichlet prior on ϕ. We chose $\mu_{R^2} \in \{0.1, 0.5\}, \varphi_{R^2} \in \{0.5, 1\}$, and $a_\pi \in \{0.5, 1\}$. The number of grouping factors K was chosen as $K \in \{0, 1\}$, implying either an R2D2 or an R2D2M2 prior. The number of covariates p was varied in $p \in \{10, 100, 300\}$. When considering the R2D2M2 prior, we used $q = p + 1$ varying terms per grouping factor K, that is, varying coefficients for each covariate plus a varying intercept. The number of levels per grouping factor was held constant with a value of $L = 20$. Finally, the residual standard deviation σ was sampled from a half Student-t with 3 degrees of freedom and scale of 1 (Gelman, 2006) and the overall intercept b_0 was sampled from a centered normal distribution with standard deviation of 5.

The covariates $x_i, i = 1, ..., p$ were sampled independently of $\tilde{\theta}$ from a multivariate normal distribution centered at the origin with a covariance matrix Σ_x, formed from a correlation matrix ρ_x which has an autoregressive structure AR(1). The correlations considered here were $\rho \in \{0, 0.5\}$. Setting $\rho = 0$ results in $\Sigma_x = I_p$ such that predictor variables are sampled independently. Finally, the outcome data $\tilde{y}$ was sampled as per Equation (1).

The obtained SBC results indicate that inference for our implementations of the R2D2 and R2D2M2 prior models is well calibrated in all considered configurations (see `https://osf.io/wgsth/` for the full results). For brevity, we only showcase a few selected results below. Specifically, we consider $(p, \mu_{R^2}, \varphi_{R^2}, a_\pi) = (100, 0.5, 1, 0.5)$ and $K \in \{0, 1\}$. The absence of grouping factors, $K = 0$, implies the use of the R2D2 prior.

Figures 7 and 8 show ECDF difference plots of several model parameters for the R2D2 and R2D2M2 models, respectively. For the R2D2 model, we show results for R^2, σ, two overall coefficients and two components of ϕ chosen randomly. For the R2D2M2 model, we additionally show two varying coefficients chosen randomly too. Selection of the presented terms can be arbitrary, since the coefficients are exchangeable in the simulation process of all configurations. If ranks were perfectly uniform, the plots would display a horizontal line, but deviations from exact uniformity are to be expected because of the simulation process. All trajectories inside the shaded blue area indicate good calibration. These results provide evidence that inference for our implementation is correctly calibrated for the quantities of interest we have investigated (i.e., all model parameters). However, as Kim et al. (2022) mention, SBC provides only a necessary

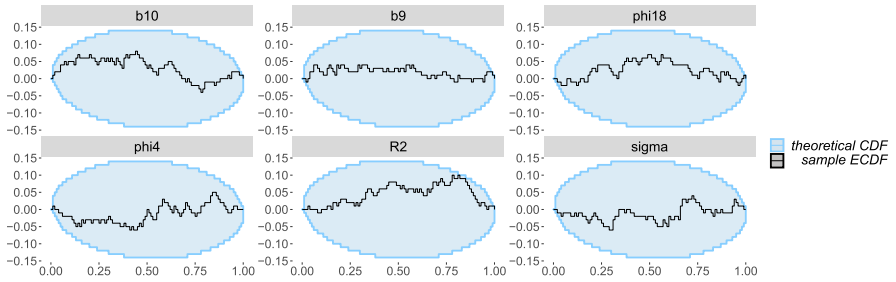


FIG 7. *ECDF difference plots for the distributions of randomly chosen parameters in the R2D2 model. The blue areas in the ECDF difference plots indicate 95%-confidence intervals under the assumptions of uniformity and thus allow for a null-hypothesis significance test of self-consistent calibration. The plots show that proper calibration has been achieved.*

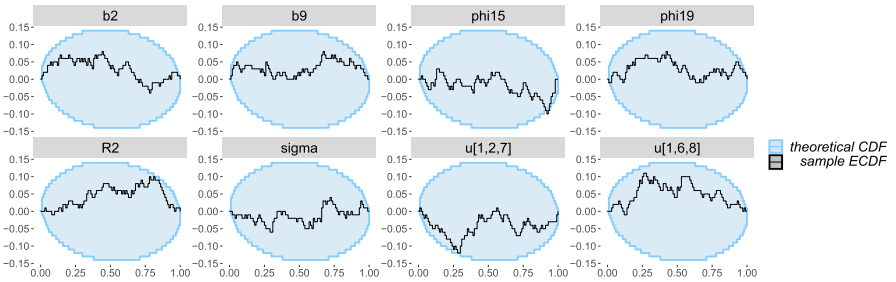


FIG 8. *ECDF difference plots for the distributions of randomly chosen parameters in the R2D2M2 model. The plots show that proper calibration has been achieved.*

but not sufficient condition for correct calibration. It is possible the incorrect calibration still occurred for some pushforward quantities of the model, although unlikely given the good calibration of all model parameters (Talts et al., 2020).

4.2. Simulations from sparse multilevel models

4.2.1. Data generation of multilevel models

We conducted simulation studies to analyze the performance of the R2D2M2 prior under different scenarios. To generate multilevel data we follow the approach of Catalina et al. (2020). Their simulation approach is sufficiently general to encompass the different forms (sparse or non-sparse) multilevel data that might be encountered in practice. A graphical model that represents the data generating process is shown in Figure 9 (see Laura Dietz (2010) for more details on this type of diagrams). The different data generating conditions are summarized in Table 6 in Appendix C.

Again, we included all possible $q = p + 1$ varying terms per grouping factor $K \in \{1, 3\}$, with $p \in \{10, 100, 300\}$. The number of levels per grouping factor

remains constant with a value of $L = 20$. The covariates $x_i, i = 1, ..., p$ are generated from a multivariate normal distribution centered at the origin with a covariance matrix Σ_x, formed from a correlation matrix ρ_x which has an autoregressive structure AR(1). The correlations considered were $\rho \in \{0, 0.5\}$. The overall intercept b_0 and the overall coefficients $b_i, i = 1, ..., p$ are simulated independently from a normal distribution with mean zero and variances $\sigma_I^2 = 4, \sigma_b^2 = 9$, respectively. For a fixed covariate i, group g and level l, the varying coefficients u_{ig_j} are simulated from a normal distribution with mean zero and variance $\sigma_g^2 = 4$.

To induce sparsity in the data generating process we set each coefficient b_i, u_{ig_j} to zero with probabilities v, z respectively. If an overall coefficient is set to zero, then we also set the corresponding varying counterparts to zero (i.e if $b_i = 0$ then $u_{ig_j} = 0, \forall g, j$). The values for v were set to $\{0.5, 0.95\}$ and z was always set equal to v. The value of the residual standard deviation σ is adjusted in each simulation to maintain the prespecified value for the true proportion of explained variance $R_0^2 \in \{0.25, 0.75\}$, to do so we used Equations (5) and (6).

4.2.2. Prior hyperparameters

In the model, we set the R^2 prior hyperparameters to $\mu_{R^2} \in \{0.1, 0.5\}$, $\varphi_{R^2} \in \{0.5, 1\}$, which leads to $a_2 \in \{0.45, 0.25, 0.9, 0.5\}$. Values for which $a_2 \leq 1/2$ allow the marginal prior distributions of the regression terms to have heavier tails than the Cauchy distribution (see Proposition 3.3) and vice-versa. The hyperparameter α is set to $\alpha = (a_\pi, ..., a_\pi)'$ with $a_\pi \in \{0.5, 1\}$ to test the different conditions near the origin, since values of $a_\pi \leq 1/2$ lead to unbounded marginal priors and vice-versa (see Proposition 3.2). Therefore, after fully crossing all these factors, we have a total of 8 different hyperparameter setups, which are summarized in Table 1. For brevity in the upcoming discussion of results, we will only refer to the indexes of these setups rather than to the specific hyperparameter values. The selected values of the hyperparameters were chosen to

TABLE 1
Configurations of the prior hyperparameters that were selected to test the R2D2M2 model under a sparse multilevel simulation setup.

Hyperparameter	1	2	3	4	5	6	7	8
a_π	0.5	1.0	0.5	1.0	0.5	1.0	0.5	1.0
μ_{R^2}	0.1	0.1	0.5	0.5	0.1	0.1	0.5	0.5
φ_{R^2}	0.5	0.5	0.5	0.5	1.0	1.0	1.0	1.0

create scenarios in which the marginal priors of the regression terms exhibit (un)boundedness near the origin and heavier (lighter) tails than the Cauchy distribution. For instance, consider setup (5) in Table 1, when $a_\pi = 0.5$ and $(\mu_{R^2}, \varphi_{R^2}) = (0.1, 1)$ (leading to $a_2 = 0.9$) we can expect that the prior will exert high shrinkage while presenting lighter tails. In this case, the prior is unbounded near the origin and there is major concentration of mass near the origin. In contrast, we can expect setup (4) to exert comparably little shrinkage, since the combination of boundedness ($a_\pi = 1$) at the origin and heavy

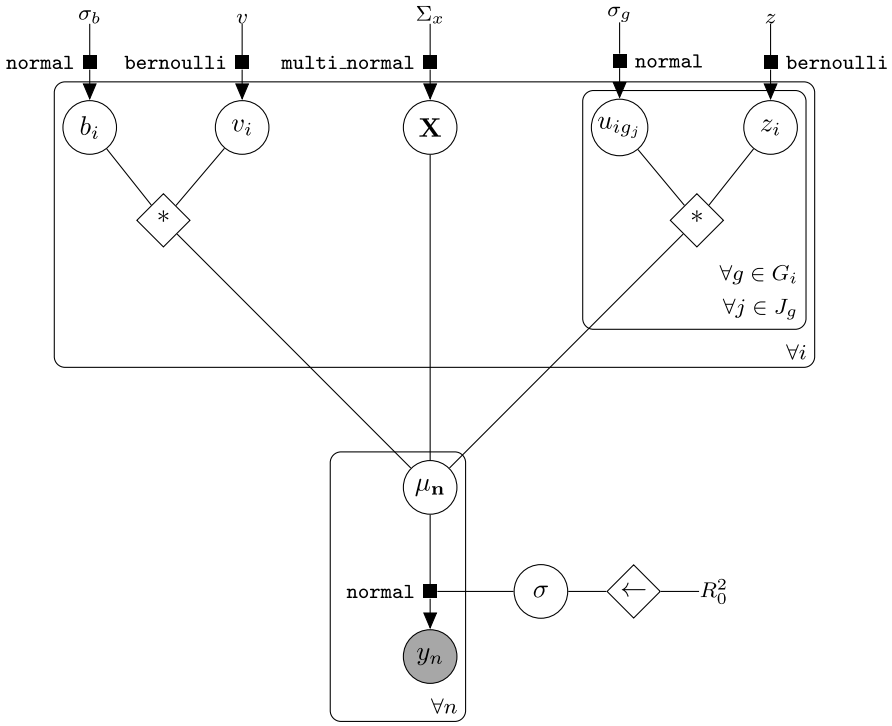


FIG 9. *A graphical model showing the data generating process for the simulations. The $*$ symbol denotes the usual product and $\leftarrow$ represents the use of Equations (5) and (6) to find the value of σ to maintain the pre-specified value of R_0^2. The normal distributions considered for the coefficients b_i, u_{ig_j}, Σ_x have an expected value of 0. The mean of y_n, given by μ_n, is calculated using Equation (1).*

tails ($a_2 = 0.25$) of the marginal priors will lead the R2D2M2 prior to shift its mass to the tails, thus focusing even more on detecting relevant signals. These different behaviors show the flexibility of the R2D2M2 prior. Similar analyses can be conducted for the other hyperparameter setups. The prior for the residual standard deviation σ was chosen as a half Student-t distribution with scale $\eta = \text{sd}(y)$ and $\nu = 3$ degrees of freedom, allowing for a finite variance but still with a heavy right tail.

The simulation setup provides an amount of 48 different data generation configurations, on which each of the different 8 hyperparameter setups will be tested. After fully crossing these conditions we have a total of 384 different simulation configurations. Table 6 in Appendix C summarizes the different conditions we have considered. For each configuration, we generated $T = 40$ datasets consisting of $N = 200$ training observations each. Out-of-sample predictive metrics were computed based on $N_{\text{test}} = 200$ test observations independent of the training data. The test set was designed to include all current grouping factors and

levels that are already in the training test, i.e., we are not performing prediction of new levels. To study the performance of the Bayesian models with the R2D2M2 prior, we have made use of several metrics. The complete list of metrics recorded for the experiments can be found in our OSF repository [1]. The primarily important ones are explained as they come up below.

4.2.3. Analysis of results

We show an analysis for three out of the total 48 different data generation configurations, whose results are representative of the overall results. They describe how the R2D2M2 prior performs in these scenarios with the different hyperparameter setups shown in Table 1. The complexity of the different scenarios considered increases in terms of number of coefficients included in the model. For the selected configurations $K = 1, 3, \rho = 0, v = z = 0.95, R_0^2 = 0.75$ and $p \in \{10, 100, 300\}$. For the different values of p, the total number of coefficients included in the linear predictor are $\{230, 2120, 6320\}$ when $K = 1$ and $\{670, 6160, 18360\}$ when $K = 3$, respectively. The $K = 1$ and $K = 3$ scenarios paint qualitatively the same picture. Of course, model performance in $K = 3$ scenarios was uniformly worse as it is considerably harder, but no new patterns (e.g., in terms of prior hyperparameter choice) emerged. For brevity, the results for $K = 3$ are only shown in Appendix C. The combination of $R_0^2 = 0.75$ and $p_i = 0.95$ implies that few (strong) signals account for most of the variability of y, a scenario which represents that only a handful of covariates have a substantial effect on the response. The R2D2M2 prior performs similarly for the other data generation setups and the results can be found on https://osf.io/wgsth/.

Estimation error We present summaries for several metrics in Table 2 considering $K = 1$. Similar results can be found when $K = 3$ in Table 7 in Appendix C. The Root Mean Square Error (RMSE) of parameter estimates (as represented by the parameter's posteriors) was calculated for all the parameters using the posterior mean as an estimator across the T replications. The RMSE is an indicator of estimation error and balances bias and variance. A low value is desirable with zero indicating the ideal case (perfectly precise and unbiased estimation). In this sense, it can be seen as criterion for proper parameter recovery. As p increases, the task of parameter recovery rises in complexity; however, RMSE values are low even in the most difficult scenarios, when $p = 300$.

Figure 10 shows the posterior shrinkage towards zero on the overall coefficients for the case in which $K = 1, p = 300$ when pooling over the different hyperparameter combinations and all simulations together. To show this, we compare the true value of the coefficient b with its posterior mean. The red diagonal line is the identity function and points that fall exactly on it indicate perfect posterior point estimation and no shrinkage. If the true b is positive (negative) and the corresponding dot is below (above) the diagonal line, then

[1] https://osf.io/wgsth/

TABLE 2
Predictive Table results with $K = 1$.

Scenario $K = 1, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200, \alpha = 0.05$									
p	Metric	1	2	3	4	5	6	7	8
10	RMSE	0.02	0.03	0.02	0.02	0.02	0.02	0.01	0.02
	elpd	−434	−427	−441	−449	−430	−433	−430	−438
	m_{eff}	32	38	33	44	31	41	31	45
100	RMSE	0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.01
	elpd	−1271	−1712	−1319	−1654	−1029	−1567	−1388	−1292
	m_{eff}	334	485	333	548	289	481	382	440
300	RMSE	0.03	0.03	0.02	0.02	0.02	0.02	0.02	0.02
	elpd	−3848	−3572	−4219	−3340	−3234	−3335	−3708	−3367
	m_{eff}	1412	1662	1560	1756	1111	1208	1330	1652

Results are shown for the Root Mean Squared Error (RMSE), expected log-pointwise predictive density (elpd) on the test set and the effective number of nonzero coefficients m_{eff}. The numbers 1-8 indicate the hyperparameter setup considered.

the posterior is shrinking the term towards zero. This 'chair-like' appearance is usually exhibited by shrinkage priors (see Carvalho et al., 2010; Van Erp et al., 2019, for more details) and is present under the different setups we have considered. This behavior indicates that the prior is exerting more shrinkage on weak signals and is consistent with the bounded influence and tail robustness properties the R2D2M2 prior possesses. Sufficiently strong signals are correctly detected and deemed as important, hence being shrunken less by the prior.

We see that, as the complexity of the problem increases (e.g., going from $K = 1$ to $K = 3$), the model responds by performing more shrinkage and by increasing the threshold a true coefficient needs to exceed to remain relatively unshrunken. This is especially clear for the $p = 300$ case since, as K increases, a flatter and wider region appears in the center of the chair-like plot, thus representing that the threshold for detection and bounded influence has increased and everything below it will be highly shrunken. This is an expected behavior of robust shrinkage priors (Carvalho et al., 2010). As complexity increases, they react by shrinking more and thus being more strict as to what they consider as a relevant signal.

Out-sample-predictive performance and shrinkage To assess predictive performance we estimate the expected log-pointwise predictive density (elpd) defined by Vehtari and Ojanen (2012) as

$$\text{elpd} = \sum_{i=1}^{N} \int p_t(\tilde{y}) \log p(\tilde{y}_i|y) d\tilde{y}_i,$$

where $p(\tilde{y}_i|y)$ is the posterior predictive distribution and $p_t(\tilde{y}_i)$ is the distribution representing the true data generating process. The elpd is as measure of predictive accuracy for the N data points taken one at a time and measures the goodness of the whole predictive distribution. We compute and report an estimator for elpd, denoted as $\widehat{\text{elpd}}$ on the test set to evaluate predictive performance. Denote the draws from the posterior distribution by $\theta^{(s)}$ where $s = 1, ..., S$, then

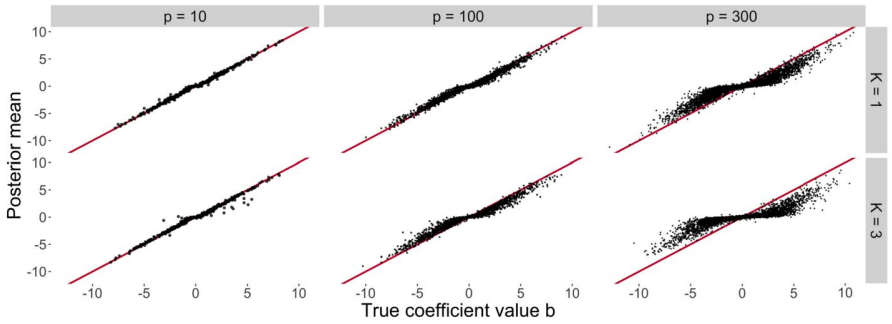


FIG 10. *Posterior shrinkage of the overall coefficients pooled over the different hyperparameter setups and simulations. We show the true value versus the posterior mean. The red diagonal line is the identity function. Shrinkage is higher for weaker (in magnitude) signals resulting in "chair-like" plots. As the complexity increases, the prior increases the threshold above which (in absolute terms) shrinkage is noticeably weaker.*

the elpd can be estimated by

$$\widehat{\text{elpd}} = \sum_{i=1}^{N} \log \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right).$$

We proceed to discuss how out-of-sample predictive performance (as measured via elpd) is intrinsically related to the amount of shrinkage done by the prior. We also discuss how prediction can be improved by imposing a higher amount of shrinkage via the hyperparameters. As a reference, the elpd values on the training set are usually close to -400 ± 50 for all cases, even as the number of covariates increases. Considering the test data data, Table 2 shows that in the comparably simple case of $p = 10$, all hyperparameter setups are providing similar point estimates for the elpd on test data and that it is similar to the elpd on training data. Figure 11 shows the distributions of the point estimator of elpd pooled across the different $T = 40$ trials arranged by setup and number of covariates along with 95% confidence intervals and the median. When $p = 10$, the distributions have similar behavior, indicating that under this relatively simple case (few predictor terms), we are able to properly generalize under the different hyperparameter setups.

When $p = 100$, Table 2 shows that there is a noticeable variability between the results of the point estimates of the elpd in the test set. Figure 11 shows that the distributions have different concentration around their median and in the width of their 95% confidence intervals. The fact that elpd decreases (while holding the number of test observations constant) is an indicator that the problem becomes more complex, to which the prior responds by increasing the amount of shrinkage it exerts.

Figure 12 shows the distributions of the posterior medians of m_{eff} across the different T trials arranged by hyperparameter setup and number of covariates, along with 95% confidence intervals and the median. Indeed, setup (5) pro-

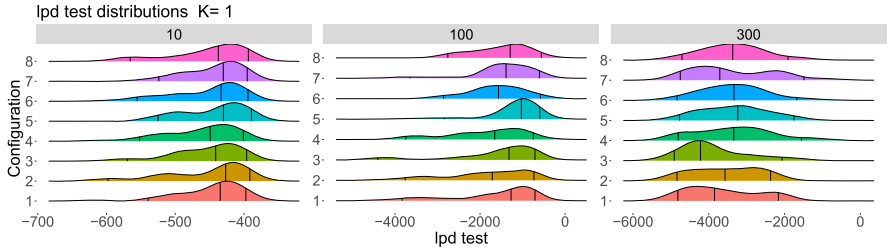


FIG 11. *Estimated densities of the elpd estimators on the test set as p increases and arranged by hyperparameter setups when* $K = 1$. *The variation depicted is across* $T = 40$ *simulation trials. The vertical lines inside each density represent the* $5\%, 50\%, 95\%$ *quantiles.*

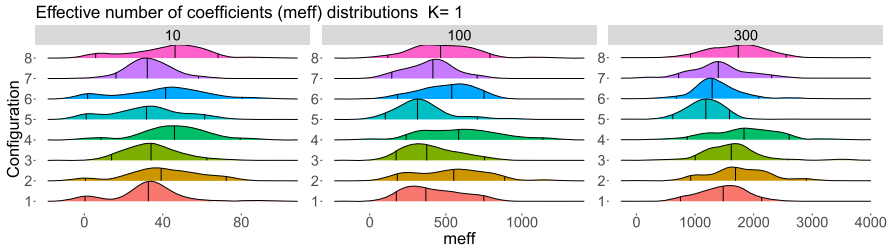


FIG 12. *Estimated densities of the posterior median of the effective number of coefficients per hyperparameter setup and number of covariates when* $K = 1$. *The variation depicted is across* $T = 40$ *simulation trials. The vertical lines inside each density represent the* $5\%, 50\%, 95\%$ *quantiles.*

duces the most shrinkage, since it sacrifices detection of large signals by having a light tail and shifting mass near the origin, thus shrinking more. It is also expected that this hyperparameter setup performs well with respect to elpd since the current simulation setup has high levels of sparsity present (around 95% of the coefficients are truly zero). A somewhat better balanced condition, which presents a trade-off between predictive performance, detection of large signals and shrinkage, would be to consider conditions similar to setup (1) where $(\mu_{R^2}, \varphi_{R^2}) = (0.1, 0.5)$, which still allows for heavy tails. Figure 13 shows the relationship between m_{eff} and elpd when estimating it over the different setups for a fixed value of p across the different T trials. Similar results are obtained when averaging over different hyperparameter setups but are not shown here. Importantly, this behavior indicates that over-shrinking has not yet occurred in any of the different configurations, as we would see an increasing and then decreasing behavior of elpd as m_{eff} increases. Overall, a general trend emerges in these results: As m_{eff} decreases, elpd increases and out-of-sample predictive performance thus improves. Therefore, to improve the latter, the user can impose hyperparameters $(\mu_{R^2}, \varphi_{R^2})$ that serve the purpose of regularizing instead of necessarily reflecting prior knowledge directly (as the latter might suggest a higher R^2 than is sensible for regularization).

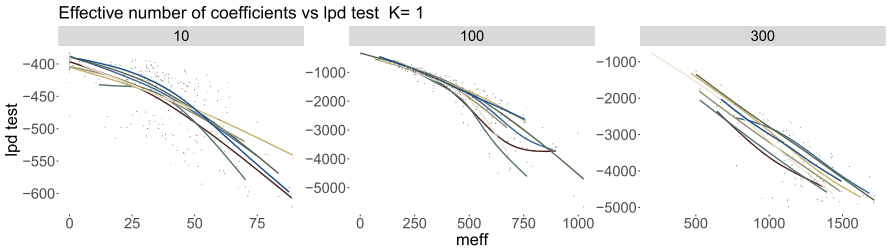


FIG 13. *Relationship between elpd and shrinkage by hyperparameter setup when $K = 1$. The non-linear relationships described by the colored lines in general shows a decreasing behavior as m_{eff} increases. Curves were estimated via thin-plate splines as implemented in the R package* `mgcv` *(Wood, 2011).*

TABLE 3
Error Table when $K = 1$ and $\alpha = 0.05$.

		Scenario $K = 1, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200, \alpha = 0.05$							
p	Metric	1	2	3	4	5	6	7	8
10	Type I error	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)
	Type II error	0.24 (0.20)	0.24 (0.17)	0.15 (0.09)	0.17 (0.12)	0.23 (0.18)	0.28 (0.23)	0.09 (0.08)	0.25 (0.17)
	FDR	0.04 (0.04)	0.01 (0.01)	0.03 (0.03)	0.01 (0.02)	0.00 (0.01)	0.02 (0.05)	0.00 (0.00)	0.09 (0.07)
100	Type I error	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)
	Type II error	0.43 (0.26)	0.40 (0.29)	0.35 (0.22)	0.45 (0.27)	0.44 (0.24)	0.47 (0.27)	0.41 (0.22)	0.36 (0.23)
	FDR	0.01 (0.01)	0.02 (0.02)	0.01 (0.01)	0.08 (0.08)	0.01 (0.01)	0.05 (0.05)	0.04 (0.04)	0.01 (0.01)
300	Type I error	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)
	Type II error	0.72 (0.58)	0.68 (0.55)	0.73 (0.57)	0.73 (0.59)	0.72 (0.59)	0.75 (0.64)	0.69 (0.58)	0.76 (0.61)
	FDR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.01)

We show the results for all the coefficients as well as results for the overall coefficients inside the parenthesis.

Credible interval coverage We evaluated the coverage properties of the R2D2M2 model by inspecting marginal credibility intervals of the regression coefficients of different size $1 - \alpha$. Table 3 reports Type I errors, Type II errors and False Discovery Rates (FDR) on the set of all coefficients regardless of the type and on the set of overall coefficients in parenthesis. To carry out the hypothesis $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ at level α, we form the credibility interval of size $1 - \alpha$ denoted by I_α. We reject H_0 at level α if $\theta_0 \notin I_\alpha$. To estimate the Type I and Type II error we consider those coefficients that are truly zero and truly non-zero, respectively. Denote by V and R the total number of hypothesis that are rejected but are true (also known as "False Positives") and the total number of hypothesis respectively (namely, "discoveries") and let $Q = V/R$, then the FDR is defined as FDR $= \mathbb{E}(Q)$ (Benjamini and Hochberg, 1995). The FDR attempts to conceptualize Type I error in the multiple comparison setting and to identify as many discoveries, while incurring in a relatively low proportion of false positives. To estimate the FDR, we consider the truly zero coefficients and make use of the same test as before, while recording Q for each simulation. When interpreting these results, keep in mind that we should not necessarily expect correct frequentist coverage from our models, since they were not built to satisfy that goal.

For a pre-specified level of $\alpha = 0.05$, Table 3 shows that Type I errors are

controlled for (at level α) in both the set of all coefficients and the set of overall coefficients (presented in parenthesis). The FDR is also controlled at this level for both types of terms. On the other hand, Type II errors increase monotonically across the hyperparameter combinations with increasing number of overall coefficients in the model. This is expected since credibility intervals contain zero even for the majority of true non-zero coefficients due to the shrinkage the prior imposes when responding to the complexity of the problem at hand. What is more, Type II errors are worse for the set of all coefficients than for the set of overall, which is the result of marginal posterior credibility intervals of varying coefficients being wider than their overall counterparts. This is bound to happen, since we have less information about the varying coefficients than we do about the overall coefficients and, as mentioned before in Section 3.2, the former are thus subject to higher shrinkage.

The size of α determines the width of the credibility intervals and therefore has a direct influence on the error rates. Since the R2D2M2 prior produces Type I errors that are sufficiently low (and much lower than α itself), we can adjust α to decrease the Type II error, sacrificing Type I error control in exchange for statistical power. Type I errors are usually known as False Positive Rates (FPR). In this context, it is usual to study the complement of the Type II error known as the True Positive Rate (TPR). Likewise, modifying α will lead to a trade-off between the FPR and TPR and allows us to study ROC curves, which are usually presented when evaluating the efficiency of a classification algorithms (Fawcett, 2006).

Figure 14 shows how we can improve the TPR while increasing the FPR as we vary $\alpha \in \{0.02, 0.05, 0.10, 0.20, 0.50, 0.66\}$ when forming the credibility interval I_α and considering all of the coefficients. Likewise, the same behavior is observed when taking into account only overall coefficients in Figure 15. To illustrate how we can make better trade-offs between TPR and FPR, consider the most complex case when $p = 300$ and a value of $\alpha = 0.2$. This results in FPR ≈ 0.2 and TPR ≈ 0.8 for the set of overall coefficients. Figures 14 and 15 also indicate that our procedure is substantially better than random classification even in the most complex cases. This occurs even without performing correction for multiple comparisons and although the prior was not built for variable selection per se. Nonetheless, we recommend to separate inference and selection procedures as argued by Piironen et al. (2020). After performing inference via the R2D2M2 prior, projection prediction methods could be applied in a second step Piironen et al. (2020); Catalina et al. (2020), but studying this is out of scope of the present paper.

Table 4 reports coverage properties of 95% credibility intervals for the set of all coefficients and for the set of all coefficients. The properties related to the subset of non-zero coefficients are reported in parenthesis. Two main points stand out: 1) Under a sparse regime, coverage of non-zero coefficients is in general a harder problem and the proportion of coverage is less for non-zero coefficients than when considering the set of all coefficients at once. Due to the high number of truly zero coefficients (95% in this scenario) the prior attempts to adapt to the sparsity at the expense of the non-zero coefficients. 2) On average, credibil-

TABLE 4
Coverage table when $K = 1$.

Scenario $K = 1, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200, \alpha = 0.05$										
p	Description	Metric	1	2	3	4	5	6	7	8
10	All	Coverage	0.99 (0.91)	0.99 (0.89)	0.99 (0.90)	0.99 (0.89)	0.99 (0.89)	0.99 (0.81)	0.99 (0.91)	0.99 (0.92)
		Width	0.86 (2.65)	1.02 (2.56)	0.96 (2.35)	1.03 (2.21)	0.85 (2.42)	0.95 (2.27)	0.91 (1.63)	0.98 (2.30)
	Overall	Coverage	0.97 (0.83)	0.98 (0.86)	0.98 (0.92)	0.98 (0.89)	0.98 (0.91)	0.96 (0.78)	0.98 (0.90)	0.96 (0.83)
		Width	0.44 (0.65)	0.52 (0.74)	0.50 (0.73)	0.52 (0.69)	0.44 (0.66)	0.50 (0.70)	0.47 (0.67)	0.50 (0.66)
100	All	Coverage	0.99 (0.71)	0.99 (0.68)	0.99 (0.66)	0.99 (0.62)	0.99 (0.68)	0.99 (0.65)	0.99 (0.70)	0.99 (0.67)
		Width	1.40 (2.32)	1.43 (1.98)	1.33 (2.25)	1.39 (1.94)	1.33 (2.35)	1.54 (2.18)	1.3 (2.22)	1.34 (1.99)
	Overall	Coverage	0.98 (0.81)	0.98 (0.80)	0.98 (0.76)	0.97 (0.77)	0.98 (0.82)	0.98 (0.8)	0.99 (0.81)	0.98 (0.72)
		Width	0.88 (1.35)	0.94 (1.36)	0.84 (1.35)	0.90 (1.33)	0.85 (1.40)	1.01 (1.45)	0.8 (1.26)	0.88 (1.33)
300	All	Coverage	0.99 (0.53)	0.99 (0.45)	0.99 (0.54)	0.99 (0.44)	0.99 (0.51)	0.99 (0.42)	0.99 (0.54)	0.99 (0.44)
		Width	1.75 (2.16)	1.66 (1.96)	1.79 (2.20)	1.68 (2.01)	1.73 (2.14)	1.78 (2.07)	1.67 (2.05)	1.76 (2.06)
	Overall	Coverage	0.98 (0.67)	0.97 (0.52)	0.98 (0.69)	0.97 (0.53)	0.98 (0.66)	0.96 (0.47)	0.98 (0.64)	0.97 (0.49)
		Width	1.41 (2.21)	1.38 (2.08)	1.44 (2.28)	1.39 (2.10)	1.4 (2.24)	1.49 (2.17)	1.34 (2.12)	1.47 (2.17)

Credibility intervals were formed at a $\alpha = 0.05$ level. We show results for the set of all coefficients mentioned in the description and for the non-zero corresponding cases inside parenthesis.

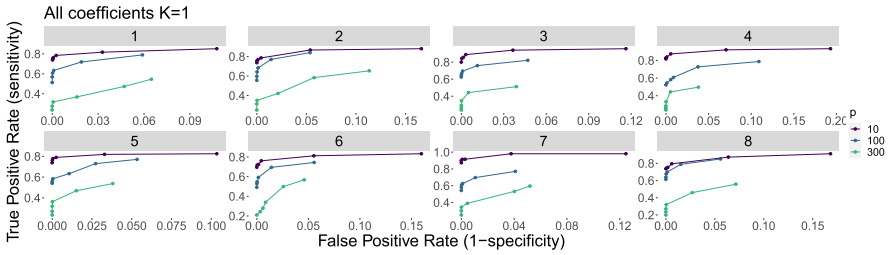


FIG 14. *ROC curves for all coefficients arranged by hyperparameter setups when $K = 1$. Points are calculated by changing α when computing credibility intervals. All ROC curves shown are above the identity line.*

ity intervals are wider for the non-zero coefficients. Since the prior adapts itself to the sparsity, it behaves conservatively when it detects a signal, performing shrinkage if the true value of the coefficient is not sufficiently high enough. Additionally, on average, the width of the credibility intervals increases with p (or with q for that matter) since there remains more epistemic uncertainty as model complexity increases given constant data set size.

4.3. Application to riboflavin production data

To illustrate the performance and usefulness of our developed models on real life challenges, we consider the `riboflavinGrouped` dataset provided by DSM (Kaiseraugst, Switzerland) and made publicly available by Buehlmann et al. (2014). This dataset has been used to study the associations between gene expressions and riboflavin (vitamin B_2) (Buehlmann et al., 2014; Lin et al., 2020). The response variable y to be predicted is the logarithm of the riboflavin production rate. The data contains the expression levels of $p = 4088$ candidate genes that might influence the riboflavin production. There are $N = 111$ observations arranged into $L = 28$ groups. Each group represents a strain (specimen)

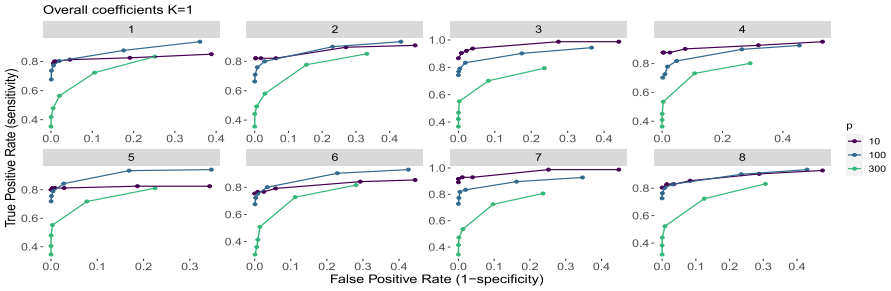


FIG 15. *ROC curves for the overall coefficients when* $K = 1$. *Points are calculated by changing* α *when computing credibility intervals. All ROC curves shown are above the identity line.*

of a genetically engineered *Bacillus subtilis* from which observations were taken. Different groups correspond to different strains of *Bacillus subtilis*, each having from 2 to 6 observations at different time points where the logarithm of the riboflavin production rate and the expression levels of the $p = 4088$ candidate genes were recorded.

The size of the data $N = 111$ is very small when compared to the number of candidate genes. Buehlmann et al. (2014) implemented a model with $p = 4088$ overall coefficients of which $q = 2$ were considered as varying among the different groups, however it is not specified in their work which are the ones they have considered to vary and why they have considered them. Lin et al. (2020) proposed modeling the association between the gene expressions and the production of riboflavin with a varying intercept model, i.e., $p = 4088$ overall terms and $q = 1$ varying term, however they do not take into account varying slopes nor models of different sizes. As an important result, both author groups reported the expression level of a specific gene, *YXLE-at*, as significantly associated with riboflavin production.

Our main goal in this case study is to test under which circumstances the R2D2M2 model is able to detect the gene *YXLE-at* as influential on the riboflavin production rate as well, even when considering much more complicated models than in earlier analysis of this dataset. To do so we built models of increasing size in terms of number of regression coefficients that will always include this main covariate of interest. A model with p overall coefficients is formed in the following way: we first include the expression level of gene *YXLE-at* as a covariate and then include the $p - 1$ covariates most correlated with *YXLE-at*; excluding those 5 covariates that are correlated with *YXLE-at* above $|r| = 0.98$, which makes them practically indistinguishable from the latter. Thus the maximal model in terms of overall coefficients is of size $p = 4083$. Once a covariate is included in the model, its varying coefficients counterpart is also included along with a varying intercept term. Including overall and varying coefficients the size of the model in terms of regression coefficients ranges from $P = 57$ when $p = 1$ to $P = 118406$ when $p = 4083$. The hyperparameters of the R2D2M2 prior are set to $(\mu_{R^2}, \varphi_{R^2}, a_\pi) = (0.1, 1, 0.25)$ to encode that we expect a fair amount of

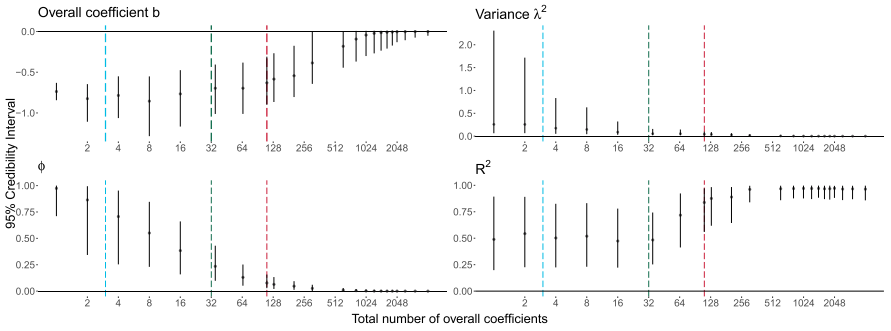


FIG 16. *Plots of posterior means and* 95% *credibility intervals related to model size in terms of the total number of overall coefficients in the model. Results for the overall coefficient and corresponding variance component of the overall coefficient of the label YXLE-at are shown in the top left and top right plots respectively. The bottom left and bottom right plot show results for the proportion of allocated variance of YXLE-at and* R^2 *respectively. The blue, green and red vertical lines indicate the point where the total number of coefficients, total number of terms and total number of overall coefficient exceed the number of observations* $N = 111$.

noise and to allow it to shrink most of the covariates to zero. This makes sense here, since only very few genes are expected to influence riboflavin production rate.

For the gene expression level of *YXLE-at*, Figure 16 shows plots of posterior means and 95% credibility intervals of the corresponding overall coefficient b, variance λ^2 of the overall coefficient, and allocation component ϕ (we drop the indices since we will only discuss the case of this gene) as well as R^2 related to the size of the model in terms of overall coefficients. The point in which the amount of observations is surpassed by the total number of coefficients $p + (q + 1)L$, number of terms $p + q + 1$ (i.e., number of components in ϕ) and number of covariates p is indicated by blue, green and red vertical lines, respectively. The gene *YXLE-at* is deemed as influential up to a certain point of p, when $p = 613$ to be precise, as the credibility intervals do not include zero; however, detection becomes more complicated as the size of the model increases, since the model responds to the complexity of the problem by shrinking coefficients stronger towards zero until this point, even the coefficients of a gene we now know to be truly relevant. This is in line with what we have shown in our simulations in Subsection 4.2 and demonstrates that the theoretically expected behavior of the prior can also be found when analysing real data. We reiterate here that shrinkage priors are not made to directly select variables but rather to provided sensible inference. If selection was really the goal, two step procedures would be recommended, of which fitting an R2D2M2 prior model would only be the first step (Piironen et al., 2020; Catalina et al., 2020).

5. Discussion

In this work, we have studied how joint regularization of all regression coefficients in Bayesian linear multilevel models can be carried out. To do so, we have have proposed the R2D2M2 prior that imposes a prior on the global coefficient of determination R^2 and propagates the resulting regularization to the individual regression coefficients via a Dirichlet Decomposition. The chosen parameterization of the R2D2M2 prior offers intuitive and interpretable hyperparameters that are easy to understand in practice and offers analysts the ability to readily incorporate prior beliefs into their multilevel model. Together with the fact that only very few hyperparameters have to be specified, this greatly simplifies prior specification for these models.

We have also derived shrinkage factors and demonstrated how to compute the effective number of non-zero coefficients of the model. This allows the analyst to have a direct understanding of the implications that the chosen hyperparameters have on the amount of shrinkage imposed by the prior. We have shown, theoretically and via simulations, that the R2D2M2 prior enables both local and global shrinkage of the regression coefficients and that it possesses vital properties that are required when working on high-dimensional regression problems, such as sufficient concentration of mass near the origin as well as heavy tails. Finally we have demonstrated, via intensive simulations and by analysing real life data, that our prior implementation is not only well calibrated, but also offers reliable estimation even in the most complex cases that we investigated.

Previously, established priors for regression coefficients in the context of multilevel modelling have only been able to either jointly regularize the overall coefficients or the varying coefficients but never both at the same time. In the case of the latter, it has been done either in the standard way of using multivariate normal distribution with fixed hyperparameters (Gelman and Hill, 2006) or by constructing a joint prior over their variance parameters (Fuglstad et al., 2019). Additionally, dealing with high dimensionality in the multilevel context has been fairly restrictive, since it has been concerned with the scenario of a larger number of overall coefficients, i.e., $p > N$, while including only very few varying terms, i.e $q \ll N$. The R2D2M2 prior allows for a joint regularization of both the overall and varying coefficients and provides the opportunity to study high dimensional scenarios in which both $p > N$ and $q > N$.

Even though it is natural to wonder how the R2D2M2 prior model compares to frequentist methods to analyze multilevel models, we stress that we are focused on very high dimensional scenarios ($p, q > N$), where we need to regularize strongly due to the complexity of the problem. Existing frequentist methods and corresponding implementations for multilevel models that we are aware of are however not able to handle a high number of varying terms ($q > N$) and also struggle with a high number of overall terms ($p > N$) due to the resulting rank-deficient design matrices. Accordingly, comparing our proposed Bayesian model with frequentist methods would not be sensible at this point.

The present work can be extended in several ways. For instance, we could consider studying regularized versions of the R2D2M2 prior as Piironen and Ve-

htari (2017) did for the original Horseshoe prior (Carvalho et al., 2010). We have shown in simulations that as the complexity of the problem increases in terms of number of coefficients included the model, the prior reacts by increasing the amount of exerted shrinkage. However, coefficients that have a sufficiently high magnitudes will basically remain unaltered. Even though this is considered a key strength of shrinkage priors by some accounts (Van Erp et al., 2019), it may not always be desirable and can be harmful when the parameters are only weakly identified by the data (Piironen and Vehtari, 2017). Alternatively, deriving joint priors for non-normal likelihoods is an important area for future research. For instance Gelman et al. (2020b) demonstrate for logistic regression that, as the number of covariates increases, quasi-complete separability becomes more likely, thus reducing the stability of the model if no substantial regularization is imposed.

Prior specification is a vital step in the Bayesian Workflow (Gelman et al., 2020b) and directly influences the performance of the models under consideration. Specifying joint priors allows us to account for the increasing complexity implied by increasing the amount of terms and parameters in the model, for example, in the shape of multilevel structure. What is more, they allow us to avoid some of the undesirable consequences of using independent, weakly informative priors in high dimensional setting. Ideally, joint priors should be able to express the user's prior knowledge while at the same time being able to perform efficient regularization, both globally and locally. These ideas together with the new developments presented here lay out a new avenue of research in the field of prior specification and elicitation.

Appendix A: Proofs

Proof of Proposition 3.1

We provide a proof for the overall coefficients $b_i, i = 1, ..., p.$. The case of the varying coefficients u_{ig_j} is handled similarly. We begin by providing an alternative parametrization of the R2D2M2 prior which is used in the proofs of propositions 3.1, 3.2, and 3.3. The case of the varying coefficients is done in a similar way.

The prior for b_i is given by

$$\tau^2 \sim \text{BetaPrime}(a_1, a_2), \phi \sim \text{Dirichlet}(\alpha), b_i|\phi, \tau^2, \sigma^2 \sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2}\phi_i\tau^2\right), \tag{23}$$

where $a_1 = \mu_{R^2}\varphi_{R^2}, a_2 = (1-\mu_{R^2})\varphi_{R^2}$. An observation from $\tau^2 \sim \text{BetaPrime}(a_1, a_2)$ can be simulated by the chain (Zhang et al. (2020))

$$\tau^2|\xi \sim \text{Gamma}(a_1, \xi), \quad \xi \sim \text{Gamma}(a_2, 1),$$

where $\text{Gamma}(a, b)$ denotes a Gamma distribution with shape a and rate b. This provides an alternative representation of the R2D2M2 prior given by

$$\begin{aligned} \xi \sim \text{Gamma}(a_2, 1), \tau^2|\xi \sim \text{Gamma}(a_1, \xi), \phi \sim \text{Dirichlet}(\alpha) \\ b_i|\phi, \tau^2, \sigma^2 \sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2}\phi_i\tau^2\right) \end{aligned} \tag{24}$$

Let α consist of a single repeating element a_π, where $a_\pi = \frac{a_1}{\dim(\alpha)} = \frac{\mu_{R^2}\varphi_{R^2}}{\dim(\alpha)}$. This is an automatic way of specifying a_π given that the user has provided a prior mean μ_{R^2} and a prior variance φ_{R^2} for R^2. Then it follows that $\phi_i\tau^2|\xi \sim \text{Gamma}(a_\pi, \xi)$. By definition, $\lambda_i^2 = \phi_i\tau^2$ and we can write the prior for b_i as

$$\xi \sim \text{Gamma}(a_2, 1), \lambda_i^2|\xi \sim \text{Gamma}\left(a_\pi, \xi\right), b_i|\sigma^2, \lambda_i^2 \sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2}\lambda_i^2\right), \tag{25}$$

or as

$$\lambda_i^2 \sim \text{BetaPrime}\left(a_\pi, a_2\right), b_i|\sigma^2, \lambda_i^2 \sim \text{Normal}\left(0, \frac{\sigma^2}{\sigma_{x_i}^2}\lambda_i^2\right).$$

Let $q_i^2 = \frac{\sigma^2}{\sigma_{x_i}^2}$, the prior marginal density of b_i is given by

$$\begin{aligned} p(b_i|\sigma) &= \int_0^\infty \frac{1}{\sqrt{2\pi q_i^2\lambda_i}^2} \exp\left\{-\frac{b_i^2}{2q_i^2\lambda_i^2}\right\} \frac{1}{\text{B}(a_\pi, a_2)} (\lambda_i^2)^{a_\pi - 1} \left(\lambda_i^2 + 1\right)^{-a_\pi - a_2} d\lambda_i^2 \\ &= \frac{1}{\sqrt{2\pi q_i^2}\text{B}(a_\pi, a_2)} \int_0^\infty \exp\left\{-\frac{b_i^2}{2q_i^2}t_i\right\} t_i^{\eta-1}(t_i + 1)^{\nu-\eta-1} dt_i \\ &= \frac{1}{\sqrt{2\pi q_i^2}\text{B}(a_\pi, a_2)}\Gamma(\eta)U(\eta, \nu, z_i), \end{aligned}$$

where $t_i = \frac{1}{\lambda_i^2}$, $\eta = a_2 + 1/2$ and $\nu = 3/2 - a_\pi$. $U(\eta, \nu, z_i)$ represents the confluent hypergeometric function of the second kind (see Jeffrey et al. (2007)) and $z_i = \frac{|b_i|^2}{2q_i^2}$. $U(\eta, \nu, z_i)$ is defined as long as the real part of η and z_i are positive, which is always the case since $a_2 > 0, q_i^2 > 0$ and $|b_i| > 0$.

Proof of Proposition 3.2

The confluent hypergeometric function of the second kind $U(\eta, \nu, z)$ satisfies the following equation (see equation 13.2.40 Olver et al. (2010))

$$U(\eta, \nu, z) = z^{1-\nu}U(\eta - \nu + 1, 2 - \nu, z). \tag{26}$$

We drop the index in z since all cases are handled equally. Substituting $\eta = a_2 + 1/2, \nu = 3/2 - a_\pi$ and $z = \frac{t^2}{2q^2}$ results in

$$U\left(a_2 + 1/2, 3/2 - a_\pi, \frac{t^2}{2q^2}\right) = (2q^2)^{1/2-a_\pi}t^{2a_\pi-1}U\left(a_2 + a_\pi, a_\pi + 1/2, \frac{t^2}{2q^2}\right).$$

The right hand side has a singularity at $t = 0$ when $a_\pi < 1/2$.

The limiting form U has as $|z| \to 0$ follows equation 13.2.18 in Olver et al. (2010) given by

$$U(\eta, \nu, z) = \frac{\Gamma(\nu - 1)}{\Gamma(\eta)} z^{1-\nu} + \frac{\Gamma(1 - \nu)}{\Gamma(\eta - \nu + 1)} + \mathcal{O}\left(z^{2-\nu}\right), \quad 1 < \nu < 2.$$

Substituting $\eta = a_2 + 1/2, \nu = 3/2 - a_\pi$ and $z = \frac{t^2}{2q^2}$ results in

$$U\left(a_2+1/2, 3/2-a_\pi, \frac{t^2}{2q^2}\right) = Ct^{2a_\pi-1} + \frac{\Gamma(a_\pi+1/2)}{\Gamma(a_2+1/2)} + \mathcal{O}\left(|t|^{1+2a_\pi}\right), \quad 0<a_\pi<1/2,$$

where $C = (2q^2)^{1/2-a_\pi} \frac{\Gamma(1/2-a_\pi)}{\Gamma(a_2+1/2)}$. The last expression is unboounded at $t = 0$ when $a_\pi < 1/2$. The case when $a_\pi = 1/2$ follows equation 13.2.19 given by

$$U(\eta, \nu, z) = -\frac{1}{\Gamma(\nu)} \left(\ln(z) + \psi(\eta) + 2\gamma\right) + \mathcal{O}\left(z \ln(z)\right), \quad 1 < \nu < 2,$$

where $\psi(\cdot), \gamma$ represent the digamma function and the Euler-Mascheroni constant respectively. Substituting $\eta = a_2 + 1/2, \nu = 3/2 - a_\pi$ and $z = \frac{t^2}{2q^2}$ gives

$$U\left(a_2 + 1/2, 3/2 - a_\pi, \frac{t^2}{2q^2}\right) = -\frac{1}{\Gamma(a_2 + 1/2)} \left(\ln\left(\frac{t^2}{2q^2}\right) + \psi(a_2 + 1/2) + 2\gamma\right) + \mathcal{O}\left(t^2 \ln\left(\frac{t^2}{2q^2}\right)\right).$$

The last expression is not defined at $t = 0$ and as $|t| \to 0$ it goes to infinity. To see the marginal priors are bounded when $a_\pi > 1/2$ it is sufficient to consider the cases given by equations 13.2.20-13.2.22 in Olver et al. (2010) and making the proper substitutions of $\eta = a_2 + 1/2, \nu = 3/2 - a_\pi$ and $z = \frac{t^2}{2q^2}$. In terms of the values a_π takes we have

$$U(\eta, \nu, t) \sim \begin{cases} \mathcal{O}\left(t^{2a_\pi - 1}\right), & 1/2 < a_\pi < 3/2 \\ \mathcal{O}\left(t^2 \ln(t^2/2)\right), & a_\pi = 3/2 \\ \mathcal{O}\left(t^2\right), & a_\pi > 3/2 \end{cases}$$

The derivative of $U(\eta, \nu, z)$ is given by equation 13.3.22 in Olver et al. (2010) as

$$\frac{d}{dz} U(\eta, \nu, z) = -\eta U(\eta + 1, \nu + 1, z).$$

Combining this with Equation (26) we have

$$\frac{d}{dz} U(\eta, \nu, z) = -\eta z^{-\nu} U(\eta - \nu + 1, 1 - \nu, z).$$

Substituting $\eta = a_2 + 1/2, \nu = 3/2 - a_\pi$ and $z = \frac{t^2}{2q^2}$ results in

$$\frac{d}{dt} U\left(a_2 + 1/2, 3/2 - a_\pi, \frac{t^2}{2q^2}\right) = Ct^{2a_\pi - 2} U\left(a_2 + a_\pi, a_\pi - 1/2, \frac{t^2}{2q^2}\right), \quad (27)$$

where $C = -(a_2 + 1/2)2^{3/2-a_\pi}(q^2)^{1/2-a_\pi}$. Equation (27) is undefined at $t = 0$ when $a_\pi < 1$.

Making $t = b_i$ or $t = u_{ig_j}$ proves the proposition.

Proof of Proposition 3.3

To prove the proposition we make use of Watson's lemma as found in Miller (2006).

Lemma A.1 (Watson's lemma)**.** *Let $0 \leq T \leq \infty$ be fixed. Assume $f(t) = t^\lambda g(t)$, where $g(t)$ has an infinite number of derivatives in the neighborhood of $t = 0$, with $g(0) \neq 0$, and $\lambda > -1$. Suppose, in addition, either that $|f(t)| < Ke^{ct}$ for any $t > 0$, where K and c are independent of t. Then it is true that for all positive x that*

$$\left| \int_0^T e^{-xt} f(t) dt \right| < \infty$$

and that the following asymptotic equivalence holds:

$$\int_0^T e^{-xt} f(t) dt \sim \sum_{n=0}^{\infty} \frac{g^{(n)}(0)\Gamma(\lambda + n + 1)}{n! x^{\lambda+n+1}},$$

for $x > 0$ as $x \to \infty$.

The marginal distribution of an overall coefficient b is given by

$$\begin{aligned} p(b|\sigma) &= \frac{1}{\sqrt{2\pi q^2}\mathrm{B}(a_\pi, a_2)} \int_0^\infty \exp\left\{-\frac{|b|^2}{2q^2}t\right\} t^{\eta-1}(t+1)^{\nu-\eta-1} dt \\ &= \int_0^\infty \exp\{-zt\} f(t) dt, \end{aligned}$$

where $z = \frac{|b|^2}{2q^2}$, $f(t) = C\, t^{\eta-1}(t+1)^{\nu-\eta-1} = t^{\eta-1}g(t)$, $C = \left(\sqrt{2\pi q^2}\mathrm{B}(a_\pi, a_2)\right)^{-1}$, and $g(t) = C\,(t+1)^{\nu-\eta-1}$. If we make $\lambda = \eta - 1$, the hypothesis $\lambda > -1$ is satisfied since $a_2 - 1/2 > -1$ for $a_2 > 0$. $g(t)$ is infinitely differentiable around $t = 0$ and $g(0) = 0$. By Watson's Lemma, since $|f(t)| < Ke^{ct}$ for all $t > 0$ where K and c are independent of t, then as $|b| \to \infty$,

$$p(z|\sigma) = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)\Gamma(\lambda + n + 1)}{n! z^{\lambda+n+1}}.$$

Truncating the sum at $n = 2$ gives

$$\begin{aligned} &p(z|\sigma) \\ &= C\left\{\frac{\Gamma(a_2+1/2)}{z^{a_2+1/2}} - \frac{(a_\pi+a_2)\Gamma(a_2+3/2)}{z^{a_2+3/2}} + \frac{(a_\pi+a_2)(a_\pi+a_2+1)\Gamma(a_2+5/2)}{z^{a_2+5/2}}\right\} \end{aligned}$$

$$+ \mathcal{O}\left(\frac{1}{z^{a_2+7/2}}\right)$$
$$\sim \mathcal{O}\left(\frac{1}{z^{a_2+1/2}}\right).$$

Therefore $p\left(|b| \,|\, \sigma\right) \sim \mathcal{O}\left(\frac{1}{|b|^{2a_2+1}}\right)$. When $a_2 < 1/2$ as $|b| \to \infty$ and comparing with $\frac{1}{b^2}$ we have that

$$\frac{p(b|\sigma)}{\frac{1}{b^2}} \sim \mathcal{O}\left(\frac{1}{b^{2a_2-1}}\right) \to \infty.$$

Proof of Proposition 3.4

We will first show that the R2D2M2 prior can be represented as Horseshoe type prior (Carvalho et al., 2010). We then proceed to make use of Theorem 2 and Theorem 3 in Carvalho et al. (2010) to show that the R2D2M2 prior with normal base distributions for the coefficients is of bounded influence.

Consider representation (25) of the R2D2M2 prior given by

$$b_i|\sigma, \lambda_i^2 \sim \text{Normal}(0, \sigma^2\lambda_i^2), \quad \lambda_i^2|\xi \sim \text{Gamma}(a_\pi, \xi), \quad \xi \sim \text{Gamma}(a_2, 1).$$

It can be shown that this representation is equivalent to

$$b_i|\sigma^2, \lambda_i^2 \sim \text{Normal}\left(0, \sigma^2\lambda_i^2\right), \quad \lambda_i^2 \sim \text{BetaPrime}\left(a_\pi, a_2\right).$$

When $a_\pi = 1/2$ and $a_2 = 1/2$, then $\lambda_i \sim \text{Cauchy}^+(0, 1)$ and we have

$$b_i|\sigma, \lambda_i^2 \sim \text{Normal}(0, \sigma^2\lambda_i^2), \quad \lambda_i \sim \text{Cauchy}^+(0, 1).$$

The Horseshoe prior (Carvalho et al., 2009) has the following representation:

$$b_i|\lambda_i, w \sim \text{Normal}(0, w^2\lambda_i^2)$$
$$\lambda_i \sim \text{Cauchy}^+(0, 1),$$

where λ_i are the local shrinkage parameters and w is the global shrinkage parameter. Therefore the R2D2M2 prior can be represented as a Horseshoe type prior where the global scale is fixed and $w = 1$. In the following consider $\sigma = 1$ and $y|b \sim \text{Normal}(b, 1)$, since by hypothesis there are no varying coefficients in the model. Theorem 2 in Carvalho et al. (2010) shows that conditioned on one sample y^*, we can write

$$\mathbb{E}(b|y^*) = y^* + \frac{d}{dy^*} \log m(y^*), \tag{28}$$

where $m(y^*)$ is the marginal density for y^* given by $m(y^*) = \int p(y^*|b)p(b)db$. Theorem 3 of Carvalho et al. (2010) shows that for the Horseshoe prior as $|y^*| \to \infty$ then

$$\lim_{|y^*|\to\infty} \frac{d}{dy^*} \log m(y^*) = 0,$$

implying that as $|y^*| \to \infty$ then $\mathbb{E}(b|y^*) \approx y^*$ and thus showing that the R2D2M2 prior is of bounded influence. The exact form of $m(y^*)$ for the R2D2 prior is given by

$$m(y^*) = \frac{1}{(2\pi^3)^{1/2}} \int_0^\infty \exp\left(-\frac{y^{*2}/2}{1+\lambda^2}\right) \frac{1}{(1+\lambda^2)^{3/2}} d\lambda.$$

Making $z = \frac{1}{\lambda^2+1}$ results in

$$\begin{aligned} m(y^*) &= \frac{1}{(2\pi^3)^{1/2}} \int_0^1 \exp\left(-1/2 y^{*2} z\right) z^{-1/2} dz \\ &= \frac{1}{\pi} \frac{\operatorname{erf}\left(y^*/\sqrt{2}\right)}{y^*}, \end{aligned}$$

where $\operatorname{erf}(\cdot)$ denotes the error function given by $\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2} dt$. Hence, $\frac{d}{dy^*} \log m(y^*)$ is given by

$$\frac{d}{dy^*} \log m(y^*) = \frac{\sqrt{\frac{2}{\pi}} e^{-y^{*2}/2}}{\operatorname{erf}\left(y^*/\sqrt{2}\right)} - \frac{1}{y^*},$$

and $\lim_{|y^*| \to \infty} \frac{d}{dy^*} \log m(y^*) = 0$. Using this and (28) we have that as $|y^*| \to \infty$ then $\mathbb{E}(b|y^*) \to y^*$.

Appendix B: Posterior computation

B.1. Stan

To obtain draws from the posterior distribution we have implemented our model in Stan (Carpenter et al., 2017; Stan Development Team, 2022), which uses a substantially extended implementation of the No-U-Turn Sampler (NUTS) from Hoffman and Gelman (2014). This sampler is an adaptive form of Hamiltonian Monte Carlo (Brooks et al., 2011). The following code is fully functional however we also maintain a current version in `https://osf.io/wgsth/`.

This implementation considers K grouping factors with multiple levels L each and D overall coefficients, including the overall intercept. All groups have the same number of levels L_g as well as the same number of varying coefficients D_g. D_g is the number of group level effects and is including the varying intercepts as well, i.e Group K level l has D_g varying coefficients. This implementation is able to work with data where D, K, L_g can vary with no need to manually modify the Stan code. However, the code can be easily modified to handle the case when groups have different amount of levels and terms per group. It is also possible to briefly modify this code in order to work with $q \leq p$ varying terms. The order of the varying terms should consider the relationship with the ith covariate since we scale by σ_{x_i}.

```
// Stan code for the R2D2M2 prior.
functions {

  vector R2D2(vector z, vector sds_X, vector phi, real tau2) {
    /* Efficient computation of the R2D2 prior
    * Args:
    *   z: standardized population-level coefficients
    *   phi: local weight parameters
    *   tau2: global scale parameter (sigma is inside tau2)
    * Returns:
      *   population-level coefficients following the R2D2 prior
    */
      return  z .* sqrt(phi * tau2) ./ sds_X ;
  }

}
data {
  int<lower=1> N;  // total number of observations
  vector[N] Y;  // response variable
  int<lower=1> D;  // number of population-level effects
  matrix[N, D] X;  // population-level design matrix
                   // including column of 1s
  int<lower=0> K; // number of groups
  vector[D-1] sds_X; // column sd of X before centering.
                    //Estimate or real values.

  //---- data for group-level effects

  int<lower=1> Lg;  // number of  levels per group (constant)
  int<lower=1> Dg; // number of coefficients per level per group
                  //(D_g constant per group)
  int<lower=1> J[N,K]; // grouping indicator matrix
                       // per observation per group K


  //---- group-level predictor values
  matrix[Dg,N] Z[K];

  //---- data for the R2D2 prior
  vector<lower=0>[ (D-1)+K+(Dg-1)*K] R2D2_alpha;
  real<lower=0> R2D2_mean_R2;  // mean of the R2 prior
  real<lower=0> R2D2_prec_R2;  // precision of the R2 prior
  int prior_only;  // should the likelihood be ignored?
}

transformed data {
  int Dc = D - 1; //
  matrix[N, Dc] Xc; //centered version of X without an intercept
  vector[Dc] means_X;//column means of X before centering
  vector[Dc] var_X;
```

```
  vector[N] Yc;
  real Ymean;
  for (i in 2:D) {
    means_X[i - 1] = mean(X[, i]);
    var_X[i-1]= sds_X[i-1]^2;
    Xc[, i - 1] = (X[, i] - means_X[i - 1]) ;
  }

  Ymean= mean(Y);
  for (i in 1:N) {
    Yc[i]= Y[i]-mean(Y);
  }

}

parameters {
  real Intercept;  // temporary intercept for centered predictors
  vector[Dc] zb; // standardized population-level effects
  matrix[Dg,Lg] z[K]; // standardized group-level effects
  real<lower=0> sigma;  // standard deviation of response

  // local parameters for the R2D2M2 prior
  simplex[Dc+K+(Dg-1)*K] R2D2_phi;
  // R2D2 shrinkage parameters
  /*Convention of indexing: First Dc for overall effects,
                             group of K for varying intercepts,
                            Batches of Dc for each grouping factor */

  real<lower=0,upper=1> R2D2_R2;  // R2 parameter

}

transformed parameters {

  vector[Dc] b;  // population-level effects

  matrix[Dg,Lg] r[K]; // actual group-level effects

  real R2D2_tau2;  // global R2D2 scale parameter
  R2D2_tau2 =  R2D2_R2 / (1 - R2D2_R2);

  // compute actual overall regression coefficients
  b = R2D2(zb, sds_X, R2D2_phi[1:Dc], (sigma^2) * R2D2_tau2);

  for(k in 1:K){
    // varying intercepts
    // Dc+k is the kth varying intercept
    r[k,1,] = (sigma * sqrt(R2D2_tau2 * R2D2_phi[Dc+k ])
              * (z[k,1,]));
    for(d in 2: Dg){
```

```
      // group level effects
      // (k-1)Dc indexes the beginning of the kth batch of scales
      r[k,d,]= sigma /(sds_X[(d-1)]) *
                sqrt(R2D2_tau2 *
                R2D2_phi[Dc+K+ (k-1)*(Dg-1) +(d-1) ]) * (z[k,d,]);

    }
  }
}

model {
  // likelihood including constants

  if (!prior_only) {
    // initialize linear predictor term
    vector[N] mu = Intercept + rep_vector(0.0, N);
    for (n in 1:N) {
      // add more terms to the linear predictor

      for(k in 1:K){
        mu[n]+=dot_product(r[k,,J[n,k]], Z[k,,n]) ;
        }

    }
    // mu+ Xc*b
    target += normal_id_glm_lpdf(Yc | Xc, mu, b, sigma);
  }


  target += beta_lpdf(R2D2_R2 | R2D2_mean_R2 * R2D2_prec_R2,
            (1 - R2D2_mean_R2) * R2D2_prec_R2); // R^2
  target += dirichlet_lpdf(R2D2_phi | R2D2_alpha); //phi

  target += normal_lpdf(Intercept | 0, 10);  // Intercept
  target += std_normal_lpdf(zb); //zb: overall effects

  for(k in 1:K){
    for(d in 1: Dg){
      target += std_normal_lpdf(z[k,d,]); // z
    }
  }

  target += student_t_lpdf(sigma | 3, 0, sd(Yc));  //

}
generated quantities {
  //---actual population-level intercept
  real b_Intercept = Ymean+Intercept - dot_product(means_X, b);

  //---y_tilde quantities of interest
```

```
  vector[N] log_lik;
  real y_tilde[N];
  vector[N] mu_tilde = rep_vector(0.0, N)+Ymean+Intercept +Xc*b;
  vector<lower=0>[(D-1)+K+(Dg-1)*K] lambdas;

  //---y_tilde calc

  for (n in 1:N) {
    for(k in 1:K){
      mu_tilde[n]+=dot_product(r[k,,J[n,k]], Z[k,,n]) ;
      }
    log_lik[n] =normal_lpdf( Y[n] | mu_tilde[n], sigma);
    y_tilde[n]=normal_rng(mu_tilde[n], sigma);
  }

  //--- lambdas

  //overall variances
  lambdas[1:Dc]= sigma^2*R2D2_phi[1:Dc]./ var_X *R2D2_tau2 ;
  //varying int variances
  lambdas[(Dc+1):(Dc+K)]= sigma^2*R2D2_phi[(Dc+1):(Dc+K)]
                          *R2D2_tau2;

  for(k in 1:K){
      // group level variances
      // (k-1)(Dg-1) indexes the start of the kth batch of scales
      lambdas[(Dc+K+(k-1)*(Dg-1)+1):(Dc+K+(k-1)*(Dg-1)+Dg-1)]=
      sigma^2*R2D2_phi[(Dc+K+(k-1)*
                    (Dg-1)+1):(Dc+K+(k-1)*(Dg-1)+Dg-1)]
                    ./ var_X*R2D2_tau2;
  }

}
```

B.2. Gibbs sampling

To illustrate a Gibbs sampling approach to obtain posterior draws we consider, without loss of generality, $K = 1$ grouping factors and consider that the first q terms (i.e covariates $x_1, ..., x_q$ with $q \leq p$) are all varying over $L = l$ levels, where we will consider that $\sigma_{xi} = 1 \forall i = 1, ..., p$. Therefore ϕ is a simplex of dimension $r = p + 1 + q$ (due to the inclusion of the varying intercept term). We consider a symmetric Dirichlet distribution for ϕ with concentration vector $\alpha = (a_\pi, ..., a_\pi)'$. Our procedure can be generalized for arbitrary concentration vectors α. For ease of notation and to be able to develop a blocked Gibbs sampling we will write model (11) in matrix form. This will drastically improve performance over moving one regression term at a time.

Denote by $b = (b_1, ..., b_p)'$, $u_j = (u_{0j}, u_{1j}, ..., u_{qj})'$ the varying coefficients in level $j = 1, ..., l$ and by $u = (u_1, ..., u_l)' \in \mathbb{R}^{l(q+1)}$ the vector containing all vary-

ing coefficients. Notice that we have suppressed the notation that indicates the group since we only have one grouping factor K. Let $y = (y_1, ..., y_N)'$ denote a vector containing the N observations $y_i, i = 1, ..., N$, X denote the standardized design matrix of dimension $N \times p$ and Z denote the varying effects matrix (see Bates et al., 2015, for details on how to construct Z). Since we only have one grouping factor, Z is a block matrix, where each block contains the observations that vary in each level. In the following, we denote the linear predictor as $\mu = Xb + Zu$.

Let $\Sigma_b = \sigma^2\Gamma_b$ denote the covariance matrix of the vector of overall coefficients b where $\Gamma_b = \text{diag}\left\{\phi_1\tau^2, ..., \phi_p\tau^2\right\}$. Let $\Sigma_u \in \mathbb{R}^{l(q+1)\times l(q+1)}$ denote the block diagonal covariance matrix of u which is given by $\Sigma_u = \sigma^2\Gamma_u$ where $\Gamma_u = \text{diag}\left\{\gamma_1, ..., \gamma_l\right\}$, and γ_j is the covariance matrix associated to u_j for $j = 1, ..., l$. In our context, $\gamma_j = \text{diag}\left\{\phi_{p+1}\tau^2, \phi_{p+2}\tau^2, ..., \phi_{p+1+q}\tau^2\right\}$ for all $j = 1, ..., l$. Finally, let I_N represent the identity matrix of order N. With the notation that has been introduced and using the representation (24), we can write the R2D2M2 model (11) in matrix form as

$$\begin{aligned} &y \,|\, \mu, \sigma^2 \sim \text{Normal}(\mu, \sigma^2 I_N), \\ &b|\sigma, \tau^2, \phi \sim \text{Normal}\left(0, \Sigma_b\right), \quad u|\sigma, \tau^2, \phi \sim \text{Normal}\left(0, \Sigma_u\right) \\ &\phi \sim \text{Dirichlet}(\alpha) \\ &\tau^2|\xi \sim \text{Gamma}(a_1, \xi), \quad \xi \sim \text{Gamma}(a_2, 1), \quad \sigma \sim p(\sigma). \end{aligned} \tag{29}$$

In the following we denote by $Z \sim \text{GenInvGaussian}\left(\chi, \rho, \nu\right)$, the Generalized Inverse Gaussian distribution (Robert, 1991) with parameters $\chi > 0, \rho > 0$ and $\nu \in \mathbb{R}$ if

$$p(z) \propto z^{\nu-1} \exp\left\{-\left(\rho z + \chi/z\right)/2\right\}.$$

We denote by $\text{InvGamma}(c, d)$ the Inverse Gamma distribution with shape c and scale d and consider that a priori $\sigma^2 \sim \text{InvGamma}(c, d)$. Consider representation (29) of the R2D2M2 prior, then the Gibbs sampling procedure is the following:

1. Set initial values for $b, u, \sigma, \phi, \tau^2, \xi$.
2. Sample $b|u, \phi, \tau, \sigma, y \sim \text{Normal}\left(\bar{b}, S_b\right)$, where
$$\begin{aligned} \bar{b} &= \left(X'X + \Gamma_b^{-1}\right)^{-1} X'(y - Zu) \\ S_b &= \sigma^2\left(X'X + \Gamma_b^{-1}\right)^{-1}. \end{aligned}$$
3. Sample $u|b, \phi, \tau, \sigma, y \sim \text{Normal}\left(\bar{u}, S_u\right)$, where
$$\begin{aligned} \bar{u} &= \left(Z'Z + \Gamma_u^{-1}\right)^{-1} Z'(y - Xb) \\ S_u &= \sigma^2\left(Z'Z + \Gamma_u^{-1}\right)^{-1}. \end{aligned}$$
4. Sample $\sigma^2|b, u, \phi, \tau^2 \sim \text{InvGamma}\left(\dot{c}, \dot{d}\right)$, where
$$\begin{aligned} \dot{c} &= c + \frac{1}{2}\left(N + p + l(q+1)\right) \\ \dot{d} &= d + \frac{1}{2}\left(||y - Xb - Zu||_2^2 + b'\Gamma_b^{-1}b + u'\Gamma_u^{-1}u\right). \end{aligned}$$

5. Sample $\tau^2|b,u,\sigma,\xi \sim \text{GenInvGaussian}\left(\chi,\rho,\nu\right)$, where

$$\chi = \frac{1}{\sigma^2}\left(b'T_b^{-1}b + u'T_u^{-1}u\right), \quad \rho = 2\xi, \quad \nu = a_1 - 1/2\left(p + l(q+1)\right),$$

where T_b and T_u are such that $\Gamma_b = T_b\tau^2$ and $\Gamma_u = T_u\tau^2$ respectively. Sampling from the Generalized Inverse Gaussian is not straightforward, but the reader can refer to Leydold and Hörmann (2011) for efficient methods.
6. Sample $\xi|\tau^2 \sim \text{Gamma}\left(a_1 + a_2, 1 + \tau^2\right)$.
7. Sample $\phi \,|\, b,u,\sigma,\xi$.
To sample $\phi \,|\, b,u,\sigma$ we make use of the following proposition

Proposition B.1. *The joint posterior of $\phi \,|\, b,u,\sigma,\xi$ has the same distribution as $(T_1/T,...,T_r/T)$ where T_j are independently drawn according to*

$$T_j \sim \textit{GenInvGaussian}\left(\frac{b_j^2}{2\sigma^2}, 2\xi, a_\pi - 1/2\right) \quad j = 1,...,p$$

$$T_j \sim \textit{GenInvGaussian}\left(\frac{\sum_{j=1}^{l} u_{ij}^2}{2\sigma^2}, 2\xi, a_\pi - l/2\right) \quad j = p+1,...,q+p+1,$$

where $r = p + 1 + q$, and $T = \sum_{i=1}^{r} T_i$.

The main idea is to integrate out τ^2 and after doing so set $\phi_i = T_i/T$. To see this works, consider the joint posterior of $\phi|b,u,\sigma,\xi$ that results from integrating τ^2,

$$\begin{aligned} p(\phi|b,u,\sigma,\xi) \propto & \prod_{i=1}^{p} \phi_i^{a_\pi - 3/2} \prod_{i=p+1}^{r} \phi_i^{a_\pi - l/2 - 1} \\ & \times \int_0^\infty (\tau^2)^{\nu-1} \exp\left\{-\frac{1}{2}\left[\rho\tau^2 + \chi/\tau^2\right]\right\} d\tau^2, \end{aligned} \tag{30}$$

where $\nu = a_1 - 1/2(p + l(q+1)), \rho = 2\xi, \chi = \frac{1}{\sigma^2}\left(\sum_{i=1}^{p}\frac{b_i^2}{\phi_i} + \sum_{i=p+1}^{r}\sum_{j=1}^{l}\frac{u_{ij}^2}{\phi_i}\right)$.
Now consider the following proposition (see Kruijer et al., 2010, Annotation (36) for the proof).

Proposition B.2. *Suppose $T_1,...,T_r$ are independent random variables, with T_i having a density f_i on $(0,\infty)$. Let $\phi_i = T_i/T$ with $T = \sum_i T_i$, then the joint density f of $(\phi_1,...,\phi_{r-1})$ supported on the simplex $\mathcal{S}^{r-1}$ has the form*

$$p(\phi_1,...,\phi_{r-1}) = \int_0^\infty t^{r-1} \prod_{i=1}^{n} f_i\left(\phi_i t\right) dt, \tag{31}$$

where $\phi_r = 1 - \sum_{i=1}^{r-1}\phi_j$.

Setting $f_i(x)$ as

$$f_i(x) \propto x^{-\delta_1} \exp\left\{\frac{b_i^2}{2\sigma^2}\frac{1}{x}\right\} \exp\left\{-\xi x\right\}, \quad i = 1, ..., p$$
$$f_i(x) \propto x^{-\delta_2} \exp\left\{-\frac{\sum_{j=1}^{l} u_{ij}^2}{2\sigma^2}\frac{1}{x}\right\} \exp\left\{-\xi x\right\}, \quad i = p+1, ..., p+q+1,$$

substituting $f_i(\phi_i\tau^2)$ into (31) and simplifying we arrive to

$$p(\phi_1, ..., \phi_{r-1}) = \prod_{j=1}^{p} \phi_j^{-\delta_1} \prod_{j=p+1}^{p+q+1} \phi_j^{-\delta_2} \int_0^\infty (\tau^2)^{r-\delta_1 p-\delta_2(q+1)-1} \times \exp\left\{-\frac{1}{2}\left[\frac{1}{\sigma^2}\left(\sum_{i=1}^{p} b_i^2/\phi_i + \sum_{i=p+1}^{r}\sum_{j=1}^{l} u_{ij}^2/\phi_i\right)/\tau^2 + 2\xi\tau^2\right]\right\} d\tau^2. \tag{32}$$

Comparing equations (30) and (32) we have $\delta_1 = 3/2 - a_\pi, \delta_2 = (l+2)/2 - a_\pi$. The proof is completed by noticing that f_j are Generalized Inverse Gaussian distributions.

h) Repeat until convergence

Appendix C: Examples

C.1. Simulation based calibration

TABLE 5
Values assigned to the hyperparameters in the Simulation Based Calibration experiment 4.1.

Description	Hyperparameter	Values
Number of simulations	T	100
Number of iterations	S	3000
Groups	K	$\{0, 1\}$
Levels	L	20
Covariates	p	$\{10, 100, 300\}$
Prior mean of R^2	μ_{R^2}	$\{0.1, 0.5\}$
Prior precision of R^2	φ_{R^2}	$\{0.5, 1\}$
Concentration parameter	a_π	$\{0.5, 1\}$
Covariance matrix of x	Σ_x	$\{I_p, AR(\rho)\}$ where $\rho \in \{0.5\}$

C.2. Simulations from sparse multilevel models

TABLE 6
Values assigned to the hyperparameters in the data generating process in experiment 4.2.

Description	Hyperparameter	Values
True value of R^2	R_0^2	$\{0.25, 0.75\}$
Groups	K	$\{1, 3\}$
Levels	L	20
Covariates	p	$\{10, 100, 300\}$
Prior mean of R^2	μ_{R^2}	$\{0.1, 0.5\}$
Prior precision of R^2	φ_{R^2}	$\{0.5, 1\}$
Concentration parameter	a_π	$\{0.5, 1\}$
Covariance matrix of x	Σ_x	$\{I_p, AR(\rho)\}$ where $\rho \in \{0.5\}$
Level of induced sparsity	v, z	$\{0.5, 0.95\}$

C.2.1. Additional tables and figures

TABLE 7
Predictive Table results with $K = 3$.

Scenario $K = 3, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200$									
p	Metric	1	2	3	4	5	6	7	8
10	RMSE	0.05	0.09	0.08	0.09	0.04	0.06	0.09	0.03
	elpd	−477	−491	−491	−461	−470	−478	−497	−499
	m_{eff}	71	93	90	80	71	115	84	88
100	RMSE	0.01	0.01	0.01	0.02	0.01	0.02	0.02	0.02
	elpd	−2700	−3049	−3806	−3440	−2899	−2920	−2744	−3681
	m_{eff}	1109	1402	1388	1764	1101	1213	1260	1629
300	RMSE	0.02	0.02	0.01	0.02	0.01	0.01	0.02	0.01
	elpd	−6210	−4413	−7064	−6285	−4480	−3413	−6471	−5501
	m_{eff}	2532	2104	2461	2722	2025	1669	2642	2405

The table shows results for the Root Mean Squared Error (RMSE), expected logpointwise predictive density (elpd) on the test set and the effective number of nonzero coefficients m_{eff}..

TABLE 8
Error Table $K = 3$.

Scenario $K = 3, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200, \alpha = 0.05$									
p	Metric	1	2	3	4	5	6	7	8
10	Type I error	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Type II error	0.34 (0.32)	0.32 (0.24)	0.23 (0.13)	0.31 (0.18)	0.23 (0.21)	0.21 (0.13)	0.17 (0.17)	0.22 (0.20)
	FDR	0.12 (0.00)	0.11 (0.00)	0.11 (0.00)	0.11 (0.00)	0.03 (0.00)	0.08 (0.00)	0.15 (0.01)	0.02 (0.00)
	TNDR	0.99 (0.95)	0.98 (0.96)	0.98 (0.98)	0.98 (0.97)	0.99 (0.96)	0.99 (0.98)	0.98 (0.97)	0.99 (0.97)
100	Type I error	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Type II error	0.50 (0.37)	0.48 (0.39)	0.40 (0.33)	0.52 (0.35)	0.35 (0.27)	0.47 (0.35)	0.49 (0.37)	0.51 (0.30)
	FDR	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.02 (0.02)	0.01 (0.00)
	TNDR	0.99 (0.97)	0.99 (0.98)	0.99 (0.98)	0.99 (0.97)	0.99 (0.98)	0.99 (0.97)	0.99 (0.97)	0.99 (0.98)
300	Type I error	0.01 (0.01)	0.01 (0.01)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	Type II error	0.83 (0.72)	0.89 (0.80)	0.85 (0.76)	0.90 (0.78)	0.82 (0.73)	0.85 (0.81)	0.87 (0.75)	0.87 (0.80)
	FDR	0.02 (0.02)	0.02 (0.02)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)	0.00 (0.00)
	TNDR	0.99 (0.96)	0.99 (0.95)	0.99 (0.96)	0.99 (0.96)	0.99 (0.96)	0.99 (0.95)	0.99 (0.95)	0.99 (0.95)

We show the results for all the coefficients as well as results for the overall coefficients inside the parenthesis. The numbers 1-8 indicate the hyperparameter setup used.

TABLE 9
Coverage Table $K = 3$.

Scenario $K = 3, \rho = 0, p_i = 0.95, R_0^2 = 0.75, N = 200, \alpha = 0.05$										
p	Description	Metric	1	2	3	4	5	6	7	8
10	All	Coverage	0.99 (0.42)	0.99 (0.46)	0.98 (0.46)	0.98 (0.43)	0.99 (0.47)	0.99 (0.51)	0.98 (0.41)	0.99 (0.60)
		Width	0.79 (1.50)	0.94 (2.16)	1.00 (2.06)	0.96 (2.04)	0.80 (2.02)	0.94 (2.14)	1.01 (1.87)	0.84 (1.84)
	Overall	Coverage	0.97 (0.80)	0.97 (0.77)	0.98 (0.90)	0.96 (0.74)	0.97 (0.80)	0.97 (0.80)	0.98 (0.88)	0.98 (0.86)
		Width	0.46 (0.75)	0.55 (0.82)	0.58 (0.89)	0.57 (0.80)	0.48 (0.73)	0.55 (0.81)	0.57 (0.84)	0.51 (0.72)
100	All	Coverage	0.99 (0.39)	0.99 (0.33)	0.99 (0.37)	0.99 (0.33)	0.99 (0.40)	0.99 (0.37)	0.99 (0.34)	0.99 (0.36)
		Width	1.07 (1.42)	0.95 (1.54)	0.96 (1.09)	1.16 (1.59)	0.96 (1.35)	1.06 (1.52)	1.01 (1.38)	1.20 (1.67)
	Overall	Coverage	0.98 (0.73)	0.97 (0.55)	0.98 (0.68)	0.97 (0.53)	0.98 (0.71)	0.97 (0.54)	0.98 (0.69)	0.97 (0.59)
		Width	0.82 (1.42)	0.75 (1.43)	0.74 (1.34)	0.91 (1.50)	0.73 (1.33)	0.85 (1.42)	0.77 (1.37)	0.95 (1.57)
300	All	Coverage	0.98 (0.29)	0.98 (0.26)	0.99 (0.29)	0.99 (0.28)	0.99 (0.30)	0.99 (0.22)	0.99 (0.28)	0.99 (0.29)
		Width	1.32 (1.55)	1.29 (1.51)	1.27 (1.55)	1.37 (1.53)	1.26 (1.58)	1.23 (1.37)	1.39 (1.61)	1.36 (1.62)
	Overall	Coverage	0.95 (0.34)	0.94 (0.23)	0.96 (0.35)	0.96 (0.24)	0.96 (0.33)	0.96 (0.25)	0.96 (0.35)	0.96 (0.24)
		Width	1.21 (2.03)	1.19 (1.79)	1.15 (1.95)	1.26 (1.91)	1.16 (2.01)	1.14 (1.73)	1.27 (2.16)	1.26 (1.92)

Credibility intervals were formed at a $\alpha = 0.05$ level. We show results for the set of all coefficients mentioned in the description and for the non-zero corresponding cases inside parenthesis. The numbers 1-8 indicate the hyperparameter setup used.

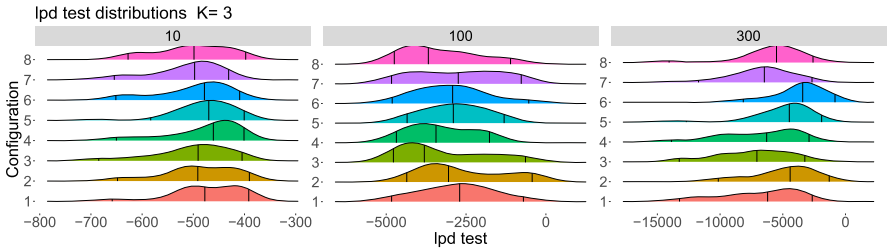


FIG 17. *Densities of the elpd estimators on the test set as p increases and arranged by hyperparameter configurations when $K = 3$. The vertical lines inside each density represent the $5\%, 50\%, 95\%$ quantiles from left to right respectively. This shows that, in average, we can expect similar results, however for single realizations differences can be observed due to the heavy tails present. Hence, proper hyperparameter specification should be done.*

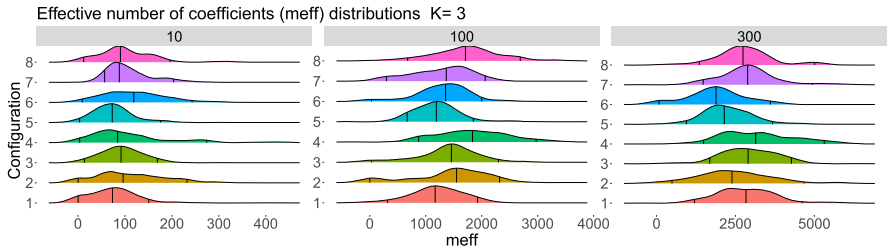


FIG 18. *Densities of the posterior median of the effective number of coefficients per hyperparameter configuration and number of covariates when $K = 3$.*

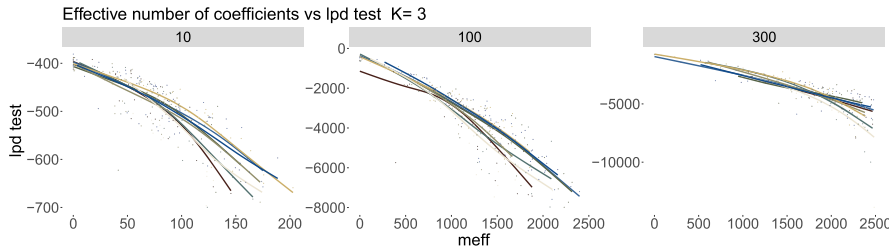


FIG 19. *Relationship between elpd and shrinkage by hyperparameter configuration when $K = 3$. The non-linear relationships described by the colored lines in general shows a decreasing behavior as m_{eff} increases.*

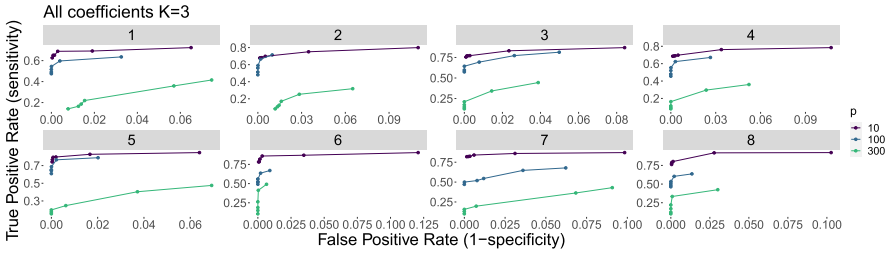


FIG 20. *ROC curves for all the coefficients arranged by hyperparameter configurations when* $K = 3$. *Points are calculated by moving* α *when forming credibility intervals. Notice that the shape of the curve is due to the fact that even for high values of* α, *False Positive Rates don't increase. All ROC curves shown are above the identity line.*

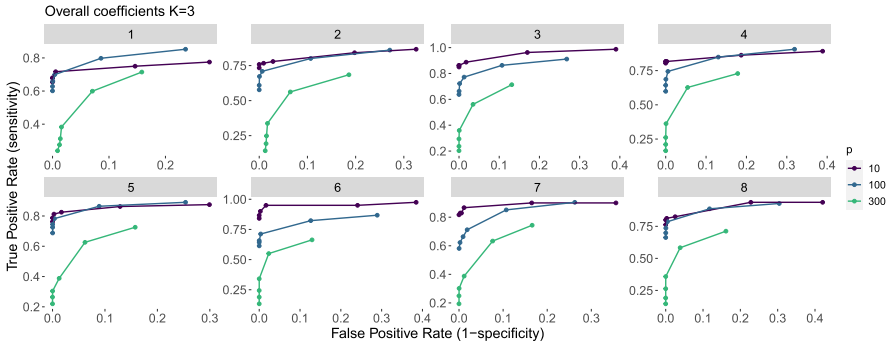


FIG 21. *ROC curves for the overall coefficients when* $K = 3$. *Points are calculated by moving* α *when forming credibility intervals. Notice that the upper bound on the curve is due to the fact that even for high values of* α, *False Positive Rates don't increase. All ROC curves shown are above the identity line.*

Acknowledgments

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2075-390740016 and DFG Project 500663361. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech).

References

Aitchison, J. and Shen, S. M. (1980). "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika*, 67(2): 261–272. URL http://www.jstor.org/stable/2335470 MR0581723

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233. URL https://doi.org/10.1080/02331888.2019.1662017 MR4034859

Barr, D. J., Levy, R., Scheepers, C., and Tily, H. J. (2013). "Random effects structure for confirmatory hypothesis testing: Keep it maximal." *Journal of memory and language*, 68 3.

Bates, D., Mächler, M., Bolker, B., and Walker, S. (2015). "Fitting Linear Mixed-Effects Models Using lme4." *Journal of Statistical Software*, 67(1): 1–48. URL https://www.jstatsoft.org/index.php/jss/article/view/v067i01

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300. URL http://www.jstor.org/stable/2346101 MR1325392

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969. URL http://www.jstor.org/stable/26363497 MR3620450

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. PMID: 27019543. URL https://doi.org/10.1080/01621459.2014.960967 MR3449048

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition.

Browne, W. J. and Draper, D. (2006). "A comparison of Bayesian and likelihood-based methods for fitting multilevel models." *Bayesian Analysis*, 1(3): 473–514. URL https://doi.org/10.1214/06-BA117 MR2221283

Buehlmann, P., Kalisch, M., and Meier, L. (2014). "High-Dimensional Statistics with a View Toward Applications in Biology." *Annual Review of Statistics and Its Application*, 1(1): 255–278. URL https://doi.org/10.1146/annurev-statistics-022513-115545

Bürkner, P.-C. (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1): 1–28. URL https://www.jstatsoft.org/index.php/jss/article/view/v080i01

Bürkner, P.-C.(2021). "Bayesian Item Response Modeling in R with brms and Stan." *Journal of Statistical Software*, 100(5): 1–54. URL https://www.jstatsoft.org/index.php/jss/article/view/v100i05

Bürkner, P.-C. and Charpentier, E. (2020). "Modelling monotonic effects of ordinal predictors in Bayesian regression models." *British Journal of Mathematical and Statistical Psychology*, 73(3): 420–451. URL https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/bmsp.12195

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1): 1–32. URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01

Carvalho, Polson, and Scott (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. URL http://www.jstor.org/stable/25734098 MR2650751

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." In van Dyk, D. and Welling, M. (eds.), *Proceedings of*

the Twelth International Conference on Artificial Intelligence and Statistics, volume 5 of *Proceedings of Machine Learning Research*, 73–80. Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA: PMLR. URL https://proceedings.mlr.press/v5/carvalho09a.html

Catalina, A., Bürkner, P.-C., and Vehtari, A. (2020). "Projection Predictive Inference for Generalized Linear and Additive Multilevel Models."

Depaoli, S. and Clifton, J. P. (2015). "A Bayesian Approach to Multilevel Structural Equation Modeling With Continuous and Dichotomous Outcomes." *Structural Equation Modeling: A Multidisciplinary Journal*, 22: 327–351. MR3357137

Fawcett, T. (2006). "An introduction to ROC analysis." *Pattern Recognition Letters*, 27(8): 861–874. ROC Analysis in Pattern Recognition. URL https://www.sciencedirect.com/science/article/pii/S016786550500303X

Fuglstad, G.-A., Simpson, D., Lindgren, F., and Rue, H. (2019). "Constructing Priors that Penalize the Complexity of Gaussian Random Fields." *Journal of the American Statistical Association*, 114(525): 445–452. URL https://doi.org/10.1080/01621459.2017.1415907 MR3941267

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. URL https://doi.org/10.1214/06-BA117A MR2221284

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. URL https://books.google.de/books?id=ZXL6AQAAQBAJ

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press. MR2262986

Gelman, A., Hill, J., and Vehtari, A. (2020a). *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press.

Gelman, A., Vehtari, A., Simpson, D., Margossian, C. C., Carpenter, B., Yao, Y., Kennedy, L., Gabry, J., Bürkner, P.-C., and Modrák, M. (2020b). "Bayesian Workflow." URL https://arxiv.org/abs/2011.01808

Good, I. J. (1962). "Theory of Probability Harold Jeffreys (Third edition, 447 + ix pp., Oxford Univ. Press, 84s.)." *Geophysical Journal International*, 6: 555–558. MR0187257

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.21.1. URL https://mc-stan.org/rstanarm

Gorinova, M. I., Gordon, A. D., and Sutton, C. (2019). "Probabilistic Programming with Densities in SlicStan: Efficient, Flexible, and Deterministic." *Proc. ACM Program. Lang.*, 3(POPL). URL https://doi.org/10.1145/3290348

Harrell, F. (2013). *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis*. Springer Series in Statistics. Springer New York. URL https://books.google.de/books?id=7D0mBQAAQBAJ

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation

for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67. URL http://www.jstor.org/stable/1267351

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15(1): 1593–1623. MR3214779

Jeffrey, A., Zwillinger, D., Gradshteyn, I., and Ryzhik, I. (2007). "8–9 - Special Functions." In *Table of Integrals, Series, and Products (Seventh Edition)*, 859–1048. Boston: Academic Press, seventh edition edition. URL https://www.sciencedirect.com/science/article/pii/B9780080471112500169 MR2360010

Jones, M. (2009). "Kumaraswamy's distribution: A beta-type distribution with some tractability advantages." *Statistical Methodology*, 6(1): 70–81. URL https://www.sciencedirect.com/science/article/pii/S1572312708000282 MR2655540

Kim, S., Moon, H., Modrák, M., and Säilynoja, T. (2022). *SBC: Simulation Based Calibration for rstan/cmdstanr models*. https://hyunjimoon.github.io/SBC/,https://github.com/hyunjimoon/SBC/

Kruijer, W., Rousseau, J., and Vaart, A. (2010). "Adaptive Bayesian Density Estimation with Location-Scale Mixtures." *Electronic Journal of Statistics*, 4. MR2735885

Kruschke, J. K. (2015). "Chapter 6 - Inferring a Binomial Probability via Exact Mathematical Analysis." In Kruschke, J. K. (ed.), *Doing Bayesian Data Analysis (Second Edition)*, 123–141. Boston: Academic Press, second edition edition. URL https://www.sciencedirect.com/science/article/pii/B9780124058880000064

Laura Dietz, B. O., Jaakko Luttinen (2010). "BayesNet." URL https://github.com/jluttine/tikz-bayesnet

Leydold, J. and Hörmann, W. (2011). "Generating generalized inverse Gaussian random variates by fast inversion." *Computational Statistics and Data Analysis*, 55(1): 213–217. URL https://www.sciencedirect.com/science/article/pii/S0167947310002847 MR2736548

Lin, J. (2016). "On The Dirichlet Distribution by Jiayu Lin."

Lin, L., Drton, M., and Shojaie, A. (2020). "Statistical Significance in High-Dimensional Linear Mixed Models." In *Proceedings of the 2020 ACM-IMS on Foundations of Data Science Conference*, FODS'20, 171–181. New York, NY, USA: Association for Computing Machinery. URL https://doi.org/10.1145/3412815.3416883

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). "Prior knowledge elicitation: The past, present, and future."

Miller, P. (2006). *Applied Asymptotic Analysis*. American Mathematical Society, 1 edition. MR2238098

Morris, M., Wheeler-Martin, K., Simpson, D. P., Mooney, S. J., Gelman, A., and DiMaggio, C. J. (2019). "Bayesian hierarchical spatial models: Implementing the Besag York Mollié model in stan." *Spatial and spatio-temporal*

epidemiology, 31: 100301.

Nakagawa, S. and Schielzeth, H. (2013). "A general and simple method for obtaining R2 from generalized linear mixed-effects models." *Methods in Ecology and Evolution*, 4(2): 133–142. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.2041-210x.2012.00261.x

Nishimura, A. and Suchard, M. A. (2022). "Shrinkage with Shrunken Shoulders: Gibbs Sampling Shrinkage Model Posteriors with Guaranteed Convergence Rates." *Bayesian Analysis*, 1–24. URL https://doi.org/10.1214/22-BA1308 MR4578057

Olver, F., of Standards, N. I., (U.S.), T., Lozier, D., Boisvert, R., and Clark, C. (2010). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press. URL https://books.google.de/books?id=3I15Ph1Qf38C MR2723248

Paananen, T., Catalina, A., Burkner, P.-C., and Vehtari, A. (2020). "Group Heterogeneity Assessment for Multilevel Models." *arXiv: Methodology*.

Park, T. and Casella, G. (2008). "The Bayesian Lasso." *Journal of the American Statistical Association*, 103(482): 681–686. URL https://doi.org/10.1198/016214508000000337 MR2524001

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155–2197. URL https://doi.org/10.1214/20-EJS1711 MR4097052

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. URL https://doi.org/10.1214/17-EJS1337SI MR3738204

Polson, N. G. and Scott, J. G. (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7(4): 887–902. URL https://doi.org/10.1214/12-BA730 MR3000018

Rights, J. D. and Sterba, S. K. (2019). "Quantifying explained variance in multilevel models: An integrative framework for defining R-squared measures." *Psychological methods*, 24 3: 309–338.

Robert, C. (1991). "Generalized inverse normal distributions." *Statistics & Probability Letters*, 11(1): 37–41. URL https://www.sciencedirect.com/science/article/pii/016771529190174P

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science*, 32(1): 1–28. URL https://doi.org/10.1214/16-STS576 MR3634300

Stan Development Team (2022). "Stan Modeling Language Users Guide and Reference Manual, Version 2.30.0." URL http://mc-stan.org/

Säilynoja, T., Bürkner, P.-C., and Vehtari, A. (2021). "Graphical Test for Discrete Uniformity and its Applications in Goodness of Fit Evaluation and Multiple Sample Comparison." MR4402179

Talts, S., Betancourt, M., Simpson, D., Vehtari, A., and Gelman, A. (2020). "Validating Bayesian Inference Algorithms with Simulation-Based Calibra-

tion."

van der Pas, S. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC.

Van Erp, S., Oberski, D., and Mulder, J. (2019). "Shrinkage priors for Bayesian penalized regression." *Journal of Mathematical Psychology*, 89: 31–50. MR3903921

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142 – 228. URL https://doi.org/10.1214/12-SS102 MR3011074

Wakefield, J. (2013). *Bayesian and Frequentist Regression Methods*. Springer Series in Statistics. Springer New York. URL https://books.google.de/books?id=OUJEAAAAQBAJ MR3025225

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman & Hall / CRC texts in statistical science. CRC Press/Taylor & Francis Group. URL https://books.google.de/books?id=0itmjwEACAAJ

Wood, S. N. (2011). "Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models." MR2797734

Yanchenko, E., Bondell, H. D., and Reich, B. J. (2021). "The R2D2 Prior for Generalized Linear Mixed Models." URL https://arxiv.org/abs/2111.10718

Zanella, G. and Roberts, G. (2021). "Multilevel Linear Models, Gibbs Samplers and Multigrid Decompositions (with Discussion)." *Bayesian Analysis*, 16(4): 1309–2770. URL https://doi.org/10.1214/20-BA1242 MR4381136

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13. URL https://doi.org/10.1080/01621459.2020.1825449 MR4436318

Article II

Aguilar, J. E., and Bürkner, P.-C. (2025). Generalized decomposition priors on R^2. *Bayesian Analysis*.

DOI: https://doi.org/10.1214/25-BA1524

The reuse of this article in the thesis is granted under the terms of the Creative Commons Attribution 4.0 International License.

Bayesian Analysis (2026) **21**, Number 3, pp. 1493–1526

Generalized Decomposition Priors on R2

Javier Enrique Aguilar [*,†,§] and Paul-Christian Bürkner[‡]

Abstract. The adoption of continuous shrinkage priors in high-dimensional linear models has gained widespread attention due to their practical and theoretical advantages. Among them, the R2D2 prior has gained popularity for its intuitive specification of the proportion of explained variance (R^2) and its theoretically grounded properties. The R2D2 prior allocates variance among regression terms through a Dirichlet decomposition. However, this approach inherently limits the dependency structure among variance components to the negative dependence modeled by the Dirichlet distribution, which is fully determined by the mean. This limitation hinders the prior's ability to capture more nuanced or positive dependency patterns that may arise in real-world data.

To address this, we propose the Generalized Decomposition R2 (GDR2) prior, which replaces the Dirichlet decomposition with the more flexible Logistic-Normal distribution and its variants. By allowing richer dependency structures, the GDR2 prior accommodates more realistic and adaptable competition among variance components, enhancing the expressiveness and applicability of R^2-based priors in practice. Through simulations and real-world benchmarks, we demonstrate that the GDR2 prior improves out-of-sample predictive performance and parameter recovery compared to the R2D2 prior. Our framework bridges the gap between flexibility in variance decomposition and practical implementation, advancing the utility of shrinkage priors in complex regression settings.

Keywords: prior specification, shrinkage priors, variance decomposition, regularization.

1 Introduction

Linear regression is one of the most widely used statistical techniques, serving as the foundation for many advanced modeling methods (Gelman and Hill, 2006; Gelman et al., 2013). However, its limitations become apparent in high-dimensional settings or when predictors exhibit multicollinearity (Hoerl and Kennard, 1970; Tibshirani, 1996; Giraud, 2014). In such cases, parameter estimation becomes unstable, and traditional models often fail to deliver reliable or interpretable results. To address these challenges, a variety of priors with exceptional theoretical and empirical properties have been developed over the last decade. These priors often assume that the true underlying model is sparse, meaning that many regression coefficients are exactly zero, and have the purpose of enforcing sparsity (Carvalho et al., 2010; van der Pas et al., 2016; van der Pas, 2021).

*Department of Statistics, TU Dortmund University, Germany, javier.aguilarr@icloud.com; url: https://jear2412.github.io

†Cluster of Excellence SimTech, University of Stuttgart, Germany

‡Department of Statistics, TU Dortmund University, Germany, paul.buerkner@gmail.com; url: https://paul-buerkner.github.io

§Corresponding author.

© 2026 International Society for Bayesian Analysis https://doi.org/10.1214/25-BA1524

Even in cases where the true model is not sparse, sparse approximations have been widely recommended and successfully applied across a range of disciplines (Tibshirani, 1996; Giraud, 2014; van der Pas et al., 2014; Hastie et al., 2015; Zhang and Bondell, 2018; Bhadra et al., 2019). Dense solutions, while theoretically possible, tend to violate the principle of parsimony by increasing model complexity, which can lead to overfitting, reduced interpretability, and higher computational costs (Giraud, 2014; Hastie et al., 2015; Yang et al., 2016). As a result, there has been significant interest in priors that balance regularization and flexibility while remaining computationally efficient.

Continuous global-local shrinkage priors are a powerful class of priors that strike a balance between sparsity and regularization while maintaining flexibility through the incorporation of both global and local shrinkage effects (Tadesse and Vannucci, 2021). Over the past decade, these priors have been extensively studied and widely adopted within the Bayesian community due to their excellent empirical performance, computational efficiency, and the availability of fast implementations (van der Pas et al., 2016; Piironen and Vehtari, 2017; van der Pas et al., 2017; Johndrow et al., 2020).

Furthermore, a broad class of continuous global-local shrinkage priors has been shown to achieve near-minimax recovery rates (Ghosal et al., 2000; van der Pas et al., 2014, 2016; Tadesse and Vannucci, 2021). General conditions over the local scales have also been established to ensure that the posterior distribution concentrates at the minimax estimation rate under these shrinkage priors (van der Pas et al., 2016; Ročková, 2018; Bhadra et al., 2019). These theoretical guarantees provide a foundation for understanding why the priors commonly used in practice perform so well: the observed empirical success is, in essence, a reflection of the strong theoretical properties underpinning these models.

Prominent examples of shrinkage priors include the horseshoe, horseshoe-plus, three-parameter beta, normal gamma, and generalized double Pareto priors (Carvalho et al., 2010; Bhadra et al., 2017; Griffin and Brown, 2010; Armagan et al., 2011, 2013). These priors typically assume the local scales are conditionally independent given the global scale. However, an alternative family of priors introduces dependencies between the scales via joint distributions, albeit in a much smaller subset of the literature. Examples include the Dirichlet-Laplace, R2D2, and its multilevel extension, the R2D2M2 prior (Bhattacharya et al., 2015; Zhang et al., 2020; Aguilar and Bürkner, 2023). These approaches model local scales as proportions using a Dirichlet distribution. Therefore the competition among variance components will only be able to gravitate towards negative dependence structures fully specified by the mean. Yet, in reality, specific coefficients or groups may compete differently for the total variability than the Dirichlet would allow for. Thus, the ability to capture more nuanced or positive dependence structures among variance components, which might better reflect real-world scenarios is severely limited.

The value of moving beyond Dirichlet distributions to more flexible alternatives has long been recognized in fields such as Compositional Data Analysis Aitchison (1986); Greenacre et al. (2023), Categorical Data Analysis (Agresti and Hitchcock, 2005), and Machine Learning in areas such as Correlated Topic Modeling (Blei and Lafferty, 2007; Chen et al., 2013). Despite their success, these methods have not been explored in the context of shrinkage priors, leaving a gap in the literature.

Among existing shrinkage priors, the R2D2 prior is particularly notable for its focus on global quantities of interest—such as the proportion of explained variance (R^2)—and its ability to jointly regularize regression coefficients (Zhang et al., 2020; Aguilar and Bürkner, 2023; Mikkola et al., 2021). By simplifying prior elicitation and enhancing interpretability, the R2D2 framework has become a valuable tool for practitioners. However, its reliance on Dirichlet decompositions limits its flexibility in modeling dependencies among variance components.

The R2D2 prior specifies a prior on R^2 and uses a Dirichlet decomposition to allocate the variance across regression terms. In this work, we address the limitations of the Dirichlet-based decomposition by introducing the Generalized Decomposition R2 (GDR2) prior, which extends the R2D2 framework using Logistic-Normal decompositions and their variants. This extension enables richer, more expressive parameterizations of dependency structures among variance components, capturing complex relationships that are inaccessible with Dirichlet-based approaches. By combining the intuitive interpretability of the R2D2 framework with increased flexibility, the GDR2 prior represents a significant step forward in the development of continuous global-local shrinkage priors.

Our work is organized as follows: we begin by discussing preliminaries and the implied distributions on the proportions of variance and R^2 by shrinkage priors in Sections 2.1, 2.2 respectively. We proceed in Section 2.3 by introducing and describing the GDR2 prior framework for high-dimensional Bayesian linear regression. To circumvent the limitations associated with the Dirichlet decomposition, we propose the use of Logistic-Normal decompositions and its variants as an alternative in the decomposition step of R^2-based shrinkage priors in Section 2.4. Section 2.5 follows with a discussion on hyperparameter specification, offering intuitive explanations for selecting hyperparameters and exploring their effects on shrinkage. We believe this contribution is particularly valuable, as hyperparameter specification for Logistic-Normal distributions is rarely addressed in the literature.

In Section 3, we present the results of simulation studies and real life experiments conducted to evaluate the capabilities of the GDR2 prior. Our findings demonstrate that allowing for richer dependency structures among variance components improves out-of-sample predictive performance and parameter recovery. For these simulations and experiments, we implemented our models in the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2024), leveraging a parameterization optimized for fast convergence and efficient Hamiltonian Monte Carlo sampling. We also provide a Slice-within-Gibbs sampler for alternative usage in the Supplementary Material (Aguilar and Bürkner, 2025). To validate our approach, we tested the GDR2 prior on three real-world benchmarks commonly used in the shrinkage prior literature. Across all scenarios, the GDR2 prior consistently led to significantly improved results. All code and replication materials are openly available on the Open Science Framework (https://osf.io/ns2cv/).

2 Methods

2.1 Preliminaries

Consider the linear regression model

$$y_n = x_n' b + \varepsilon_n,\ n = 1, \ldots, N, \tag{1}$$

where y_n is the nth response value, x_n is the K dimensional vector of covariates for the nth observation, $b = (b_1, \ldots, b_K)'$ is the vector of regression coefficients, and $\varepsilon_n \sim (0, \sigma^2)$ is the residual error, with σ being the residual standard deviation. Continuous Global-Local (GL) shrinkage priors (van der Pas et al., 2016; Van Erp et al., 2019; van der Pas, 2021) are a special type of continuous prior distributions that arise from scale mixtures of normals (West, 1987). They take on the form

$$b_k \mid \lambda_k^2, \tau^2, \sigma^2 \sim \left(0, \sigma^2 \lambda_k^2 \tau^2\right),\ \lambda_k \sim \pi(\lambda_k), k = 1, \ldots, K,\ \tau \sim \pi(\tau), \sigma \sim \pi(\sigma), \tag{2}$$

where λ_k represents local scales unique to each regression coefficient b_k, τ denotes a global scale shared across all coefficients, and $\pi(\cdot)$ represents the corresponding (hyper) prior distributions. Priors on b of the form (2) centered at zero are designed to shrink each coefficient towards zero thus favouring more sparse solutions. These priors have demonstrated empirical success among practitioners, and their theoretical foundations justify their widespread application (Ghosh et al., 2019; Johndrow et al., 2020; Follett and Yu, 2019; Kohns and Szendrei, 2024; Tadesse and Vannucci, 2021).

The choice of $\pi(\lambda_k)$ and $\pi(\tau)$ produces different GL shrinkage priors and will have an important effect on their properties. Popular shrinkage priors include the Horseshoe (Carvalho et al., 2010), the Normal-Gamma (Griffin and Brown, 2010), Generalized Double Pareto (Armagan et al., 2013), the Dirichlet-Laplace (Bhattacharya et al., 2015), the Regularized Horseshoe (Piironen and Vehtari, 2017), and the R2D2 prior (Zhang et al., 2020). The global scale τ controls the overall sparsity, ideally represent the proportion of true signals (Piironen and Vehtari, 2017; Van Erp et al., 2019; van der Pas, 2021). The local scales λ_k can either counteract or enforce the shrinkage towards zero.

The amount of shrinkage exerted on each coefficient towards zero can be quantified by studying the posterior distribution of the coefficients b given the scales λ_k, τ, σ and the observations y. The conditional posterior of the regression coefficients b is $b \mid y, \lambda, \tau, \sigma \sim \left(\bar{b}, \Sigma_b\right)$, with mean $\bar{b} = \quad (b \mid y, \lambda, \tau, \sigma) = \left(X'X + 1/\tau^2 \Lambda^{-1}\right)^{-1} X'y$ and covariance matrix $\Sigma_b = \sigma^2 \left(X'X + 1/\tau^2 \Lambda^{-1}\right)^{-1}$, where $\Lambda = \{\lambda_1^2, \ldots, \lambda_K^2\}$ is the diagonal matrix containing the local scales. If X is of full rank, then the conditional posterior mean can be expressed as $\bar{b} = \tau^2 \Lambda \left(\tau^2 \Lambda + (X'X)^{-1}\right) \hat{b}$ where $\hat{b}$ is the Maximum Likelihood Estimator (MLE). This representation highlights that the conditional posterior mean is a shrunken version of the usual MLE estimate.

For illustration, consider $X = I$, i.e. the normal means problem (Stein, 1981; van der Pas et al., 2016; Bhadra et al., 2017) and let $\sigma = 1$. In this case $\bar{b}_k = \left(1 - \frac{1}{1+\lambda_k^2 \tau^2}\right) \hat{b}_k = (1 - \kappa_k) y_k$, where $\kappa_k := \frac{1}{1+\lambda_k^2 \tau^2}$. The quantity κ_k serves as a *shrinkage factor* and

allows us to assess how much shrinkage the prior exerts on the maximum likelihood estimator of b_k (Polson et al., 2013; Bai and Ghosh, 2019; Bhadra et al., 2019; Aguilar and Bürkner, 2023). Importantly, by an application of Fubini's theorem, we can show that $\mathbb{E}[b_k \mid y_k] = (1 - \mathbb{E}[\kappa_k \mid y_k])\, y_k$, demonstrating that the posterior mean of b_k is at most y_k (Carvalho et al., 2010; Efron, 2011).

The prior on κ_k is determined by the interplay between the priors on λ_k and τ. Thus, calibrating shrinkage requires careful consideration of these priors. Notably, for a fixed value of τ, the definition of κ_k reveals that the posterior mean of b_k is influenced by the value of the local scale λ_k. Small values of λ_k will shrink b_k towards zero, whereas large values push the posterior mean of b_k towards the observation y_k. If either $\tau \to \infty$ or $\lambda_k \to \infty$ then the MLE is recovered.

In general, a good shrinkage prior should possess the following characteristics (Castillo and van der Vaart, 2012; Piironen and Vehtari, 2017; Van Erp et al., 2019; Bhadra et al., 2019; van der Pas, 2021): 1) *Heavy tails:* Sufficient mass in the tails of the prior is crucial to properly recover signals (i.e., truly nonzero coefficients). 2) *Sufficient mass near zero:* Shrinkage priors should allocate enough prior mass near zero in order to shrink redundant (truly zero) coefficients towards it. 3) *Efficient and stable sampling:* While theoretical properties are important, designing shrinkage priors should also consider efficient sampling from the posterior distribution.

Continuous GL shrinkage priors have gained popularity due to their capability to discern noise from signals, while yielding solutions that avoid a search over the entire space of models. They effectively shrink the influence of covariates considered unimportant towards zero while recovering the values of signals. This holds true even in scenarios where the true vector of regression coefficients b is ultra sparse (Bhadra et al., 2017; Song and Liang, 2017; van der Pas et al., 2017). Since continuous GL shrinkage priors cannot product *exact* zero estimates, additional posterior variable selection is required to induce sparsity in the posterior estimates, leading to a two step procedure: first perform inference of the GL shrunken model and second employ a decision rule to perform variable selection (Bai and Ghosh, 2019; Piironen et al., 2020; van der Pas et al., 2017; Zhang and Bondell, 2018; Pavone et al., 2020).

2.2 Implied Priors by the Local and Global Scales

Consider the linear regression model (1), where the covariates are standardized, such that $\mathbb{E}(x) = 0$ and $\mathrm{var}(x) = \Sigma_X$, with Σ_X having a diagonal of ones. Assume a prior distribution for the regression coefficients b such that $\mathbb{E}[b] = 0$ and $\mathrm{var}(b) = \sigma^2\Lambda$, where Λ is a diagonal matrix with entries $\lambda_1^2, \ldots, \lambda_K^2$. The conditional variance of the linear predictor $x'b$ is given by (Gelman et al., 2020):

$$\mathrm{var}(x'b) = \sigma^2 \sum_{k=1}^{K} \lambda_k^2. \tag{3}$$

We define the quantity $\omega^2 := \sum_{k=1}^{K} \lambda_k^2$ as the *total variance*. Equation (3) provides insight into the prior distribution for the proportion of explained variance, R^2 (Gelman

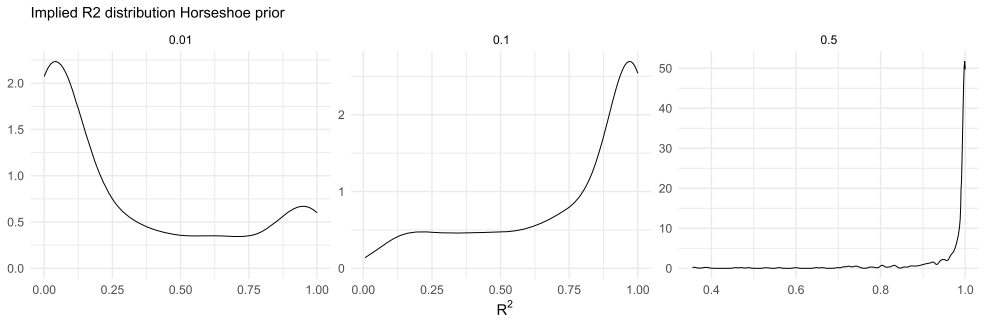


Figure 1: Shrinkage prior implied distribution for R^2 when considering the Horseshoe prior for different values of τ, which is interpreted as the a priori proportion of signals.

et al., 2020), induced by a GL shrinkage prior. The proportion of explained variance is defined as the square of the correlation coefficient between y and the linear predictor $x'b$:

$$R^2 := \text{corr}^2(y, x'b) = \frac{\text{var}(x'b)}{\text{var}(x'b) + \sigma^2} = \frac{\omega^2}{\omega^2 + 1}. \tag{4}$$

Thus, there is a one-to-one relationship between the priors set on R^2 and those implied on ω^2. Specifying a prior for b and σ implicitly defines a distribution for R^2. For instance, when b_k and σ have weakly informative priors (Gelman, 2006a), as shown in Aguilar and Bürkner (2023), the resulting prior for R^2 is highly concentrated near 1, even for moderate values of K. To illustrate the case of shrinkage priors, assume $X = I$ and use the well-known Horseshoe prior (Carvalho et al., 2010), which is specified as:

$$b_k \mid \lambda_k \sim \quad (0, \lambda_k^2), \;\; \lambda_k \mid \tau \sim C^+(0, \tau), \tau \mid \sigma \sim C^+(0, \sigma), \tag{5}$$

where $C^+(0, \tau)$ denotes a Half Cauchy distribution (Gelman, 2006a; Polson and Scott, 2012) with scale parameter τ. We assume a fixed value for τ as a user-specified hyperparameter, rather than assigning it its own prior. As demonstrated by van der Pas et al. (2016), it is crucial for τ to reflect the proportion of signals in order to guarantee signal recovery.

Figure 1 shows the implied distribution of R^2 when using the Horseshoe prior with different values for τ. Interpreting τ as the proportion of true signals, we observe that for very low τ values, the implied distribution of R^2 tends to concentrate near zero. On the other hand, when a user believes that even a small proportion of the elements in b are nonzero, i.e., $\tau \approx 0.1$, the distribution shifts toward one, indicating that the Horseshoe prior is focusing on identifying signals.

A significant issue arises when, despite having a low prior expectation for the proportion of nonzero coefficients, the model predicts high values of R^2. This leads to overestimation of the importance and magnitude of potential signals, which can conflict with user intuition about how the number and strength of signals relate to the

proportion of explained variance. Such an overestimation could mislead the user into thinking the model explains more variance than it realistically does, which could skew model interpretation and decision-making. Moreover, controlling the properties of the resulting distribution over R^2, such as its mean and variance, becomes challenging, as they are implicitly determined by the prior and the data. Given the total variance ω^2 in the model, we can compute the *proportion of explained variability* for each coefficient, defined as

$$_k \coloneqq \frac{\lambda_k^2}{\sum_{k=1}^{K} \lambda_k^2}. \tag{6}$$

Since $_k \geq 0$ for all $k = 1, \ldots, K$ and $\sum_{k=1}^{K} \ _k = 1$, the vector $\coloneqq (\ _1, \ldots, \ _K)$ lies in the $K - 1$ dimensional simplex. Thus, the joint distribution of the λ_k values induces a distribution on the simplex, represented stochastically by . Understanding the implied distribution of provides insights into how the total variance ω^2 is distributed among the coefficients and how they compete for it. This understanding can help determine how much of the total variance is allocated to each coefficient, which is key for assessing model fit and identifying influential predictors. For example, if we assume $\lambda_k \sim \text{Gamma}(\alpha_k, 1)$, we obtain the Dirichlet distribution $\sim \text{Dirichlet}(\alpha)$, where $\alpha \coloneqq (\alpha_1, \ldots, \alpha_K)$ is the concentration vector (Lin, 2016). In this case, the correlations between the elements of are negative and determined by the mean of the distribution. As with the implicit distribution for R^2, specifying a prior on λ does not necessarily give us direct control over the properties of the implied distribution for , except in certain special cases, such as the one just mentioned.

The emphasis on implied quantities that are more intuitive and user-friendly has largely been overlooked in the shrinkage prior literature, with only a few notable exceptions (Zhang and Bondell, 2018; Zhang et al., 2020; Aguilar and Bürkner, 2023). This oversight is understandable, as statisticians have primarily concentrated on developing robust automatic procedures capable of tackling complex tasks. However, we believe that emphasizing the more intuitive perspective could not only enhance user comprehension and the practical applicability of well-established shrinkage priors but also inspire the development of new, innovative methodologies.

2.3 Generalized Decomposition R2 Priors

We have discussed how the specification of a shrinkage prior on the regression coefficients b implies priors for interpretable quantities such as R^2 and . It is equally insightful to explore the reverse scenario, that is, establishing a prior over the proportion of explained variance R^2 and the proportions of total variance to derive a prior over the regression coefficients b. This concept has been explored before to define prior distributions in the context of additive regression models of varying complexity (Zhang et al., 2020; Aguilar and Bürkner, 2023). However, they always considered $\sim \text{Dirichlet}(\cdot)$. While this choice is certainly a reasonable one, it raises the question of what happens when follows a distribution other than Dirichlet.

Employing alternative distributions for allow us to probe how the proportions of total variance $_k, k = 1, \ldots, K$ compete for the total variability ω^2. Left unattended

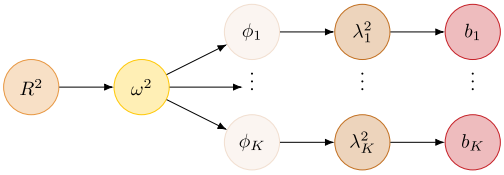


Figure 2: Schematic of the GDR2 prior construction. First, a distribution is assigned to R^2, then its uncertainty is propagated to the regression terms via the simplex distribution of the proportions of explained variance. Finally, a distribution is used to allocate variance to each regression term.

and without any proper control, this competition is naturally guided towards a certain degree of negative dependence structures, i.e., if $_k$ increases then $_j, j \neq k$ decreases since there is competition for a total quantity. Specifically, the Dirichlet distribution shows this phenomenon (Aitchison, 1986; Kruijer et al., 2010; Lin, 2016). However, in reality, specific (groups of) coefficients may compete differently for the total variability than the Dirichlet would assume and allow for. Indeed, as we show in our simulations and real-world case studies, other prior choices may lead to more favourable results.

To this end, we introduce a prior distribution family that provides greater flexibility in capturing the intricate relationships among the proportions of explained variance. We call them *Generalized Decomposition R2* (GDR2) priors. The main idea is schematically presented in Figure 2. We begin by setting a prior over R^2. According to Equation (4), there exists a one to one relationship between R^2 and the total variability ω^2, immediately establishing a prior distribution on ω^2. Subsequently, we select a distribution for the proportions of explained variance that can properly represent a desired dependence structure. Afterwards, we shape the variances of the regression coefficients b_k by setting $\lambda_k^2 = \ _k\omega^2, i = 1, \ldots, K$. Finally, we let b_k follow a normal distribution centered at zero with variance $\sigma^2\lambda_k^2$.

We opt for a Beta distribution to characterize R^2 and we write $R^2 \sim \text{Beta}(\cdot, \cdot)$. The Beta distribution, offering various parametrizations (Kruschke, 2015), is chosen for its flexibility. We specifically parametrize it in terms of the prior mean μ_{R^2} and prior "precision" $_{R^2}$ (also known as mean-sample size parameterization). This choice facilitates an intuitive integration of prior knowledge into the R^2 prior. The hyperparameters μ_{R^2}, $_{R^2}$ are interpretable expressions of domain knowledge that represent the existing relationship between the included covariates and the response variable.

If (a_1, a_2) are the canonical shape parameters of the Beta distribution, then $\mu_{R^2} = \frac{a_1}{a_1+a_2}$ and $_{R^2} = a_1 + a_2$. This selection implies that ω^2 follows a Beta-Prime distribution, with parameters μ_{R^2}, $_{R^2}$, which we denote as $\omega^2 \sim \text{BetaPrime}(\mu_{R^2}, \ _{R^2})$. Figure 3 illustrates the flexibility that the Beta distribution can offer in expressing prior knowledge about R^2. The figure also depicts the corresponding Beta Prime prior for ω^2. For instance, for values of $(\mu_{R^2}, \ _{R^2}) = (0.5, 1)$, we obtain a bathtub-shaped prior on R^2, concentrating the majority of the mass near the extremes $R^2 = 0$ and $R^2 = 1$. This

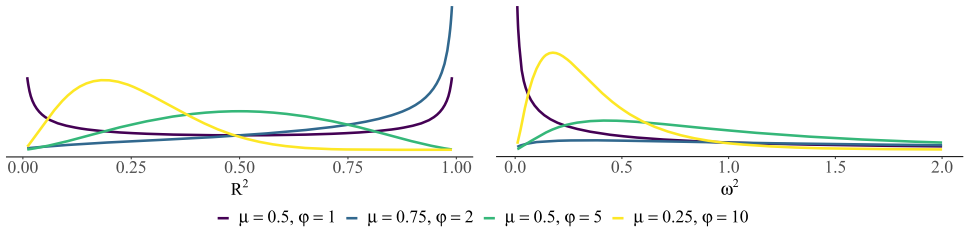


Figure 3: Beta and Beta Prime densities for various values of the prior mean μ_{R^2} and precision $_{R^2}$.

prior behavior signals the user's expectation of a model containing either a substantial amount of either noise or signal, respectively.

Moving forward, the *(probability) simplex* of dimension $K-1$ ($K \in$), is defined as $^K := \{x \in {}^K | \sum_{k=1}^K x_k = 1, x_k \geq 0, k = 1, \ldots, K\}$. Let $p(\ | \ \nu)$ represent the chosen distribution for on the simplex ^{k}, where ν represents specific hyperparameters related to the chosen distribution. For example, in the case of the Dirichlet distribution $\nu = \{\alpha\}$. Although the Dirichlet distribution aids in providing analytical results, interpretation, computational efficiency, and alignment with common user knowledge, it exhibits undesirable properties in practice, which we discuss further in Section 2.4.

The next step is to specify a distribution for the coefficients that captures the attributed variance for each regression coefficient b_k, given by $\lambda_k^2 = {}_k\omega^2$ for $k = 1, \ldots, K$. To do so, we define $b_k|\sigma, \lambda_k \sim (0, \sigma^2\lambda_k^2)$, which integrates our prior into the GL shrinkage prior framework, typically modeled as a scale mixture of normals (West, 1987). In contrast, Bhattacharya et al. (2015) and Zhang et al. (2020) have used double exponential distributions for b_k to express the attributed variances, although the former did not account for the perspective of variance proportions. Additionally, Aguilar and Bürkner (2023) adopted normal distributions to model the attributed variances, offering a different approach to the distribution of the coefficients.

If model (1) includes an intercept b_0, the prior specification is simplified by setting a prior on the centered intercept $\tilde{b}_0$, which is implied when $(x_i) = 0$. The original intercept b_0 can then be recovered after model fitting using a simple linear transformation (Bürkner, 2017; Goodrich et al., 2020). Common choices for priors on b_0 include a normal prior with mean (y) and a user-chosen scale, which depends on the scale of y, or Jeffrey's prior, which is improper and flat in this case (Good, 1962). Next, we specify a prior for the residual variance σ^2 (or equivalently the residual standard deviation σ). Following the recommendations of Gelman (2006b), we set a Half Student-t prior on σ with ν degrees of freedom and scale η. Consistent with Bürkner (2017) and Aguilar and Bürkner (2023), we set $\eta \approx \text{sd}(y)$, as both the prior's expected mean and variance are

proportional to η. Together, the full GDR2 model is given by

$$\begin{aligned} y_n \sim \quad & (\mu_n, \sigma^2), \quad \mu_n = b_0 + \sum_{k=1}^{K} x_{nk} b_k, \quad n = 1, \ldots, N, \\ b_0 \sim p(\cdot), \quad b_k \sim \quad & (0, \sigma^2 \ {}_k\omega^2), \quad k = 1, \ldots, K, \\ \omega^2 = \frac{R^2}{1 - R^2}, \quad & \sim p(\nu \ ), \quad R^2 \sim \text{Beta}(\mu_{R^2}, \ {}_{R^2}), \quad \sigma \sim p(.). \end{aligned} \tag{7}$$

While, in this paper, we focus on the linear regression case, the additive structure of the prior readily allows its extension to linear multilevel models or other kinds of additive models if desired.

2.4 Distributions on the Simplex

In this section, we explore potential distributions for defined on the simplex. While extensions of the Dirichlet distribution are often considered, we do not pursue them here, as they share similar limitations (Connor and Mosimann, 1969; Barndorff-Nielsen and Jørgensen, 1991; Ongaro and Migliorati, 2013; Chow, 2022). Instead, we draw on methods from Compositional Data Analysis (CDA) (Aitchison, 1986; Boogaart and Tolosana-Delgado, 2013; Greenacre et al., 2023), Categorical Data Analysis (CatDA)(Agresti and Hitchcock, 2005) and Correlated Topic Modeling (CTM) (Blei and Lafferty, 2007; Chen et al., 2013), where the logistic normal distribution (and its variants) is the preferred approach for analyzing simplicial variables.

The Dirichlet Distribution

The Dirichlet distribution is widely employed for modeling data existing in the simplex (Gupta and Richards, 2001; Lin, 2016). A vector $= (\ _1, \ldots, \ _K)' \in \ ^K$ follows a Dirichlet distribution with concentration vector $\alpha := (\alpha_1, \ldots, \alpha_K)', \alpha_k > 0, k = 1, \ldots, K$ if its density has the following form:

$$p(\ \ |\alpha) = \frac{\Gamma(\alpha_+)}{\prod_{k=1}^{K} \Gamma(\alpha_k)} \prod_{k=1}^{K} \ {}_k^{\alpha_k - 1} \mathbb{1}(\ \in \ ^k), \tag{8}$$

where $\alpha_+ = \sum_{k=1}^{K} \alpha_k$. The mean and covariance of the distribution are respectively expressed as: $(\ _k) = \frac{\alpha_k}{\alpha_+} =: \tilde{\alpha}_k$ and $\text{cov}(\ _k, \ _j) = \frac{\delta_{kj}\tilde{\alpha}_k - \tilde{\alpha}_k\tilde{\alpha}_j}{1+\alpha_+}$, where δ_{kj} is the Kronecker delta function. If $k \neq j$, then $\text{cov}(\ _k, \ _j) = \frac{-\tilde{\alpha}_k\tilde{\alpha}_j}{\alpha_+ + 1} < 0$ for all k, j. Once the mean vector $\tilde{\alpha}$ is chosen, only the scalar α_+ remains to express the entire variance-covariance structure. In particular, the covariance function is always proportional to the product of the corresponding means.

The Dirichlet distribution gained prominence during the conjugate prior era, as it serves as the conjugate prior for the multinomial likelihood (Ongaro and Migliorati, 2013; Lin, 2016; Wang and Polson, 2024). Its interpretability is another advantage, the

hyperparameters α_k can easily be interpreted in relation to the distribution's behavior. In particular, the quantity $\tilde{\alpha}_k = \quad (\quad _k)$ represents the relative a priori importance of the kth element $\quad _k$. To illustrate, consider $\alpha = (1, \ldots, 1)'$, which gives a distribution that is flat over all possible simplexes, leading to a suitable choice in the absence of additional prior knowledge. Generally, setting $\alpha = (a_\pi, \ldots, a_\pi)'$ with a concentration parameter $a_\pi > 0$ (i.e., a *symmetric Dirichlet distribution*) drastically reduces the number of hyperparameters to specify and allows us to globally control the shape of the Dirichlet distribution with a single value. A symmetric Dirichlet distribution is useful when there is no prior preference that favors one component over another. Values $a_\pi < 1$ concentrate mass on the edges of the simplex, while $a_\pi > 1$ results in concentration around the center of the simplex. The user can also specify asymmetric Dirichlet distributions by choosing different values for the individual α_k. This represents different a priori expected importance of the corresponding components.

A major limitation of the Dirichlet distribution is its inherent lack of flexibility in modeling dependencies among the elements of $\quad$. Its covariance structure is determined entirely by the mean, restricting the ability to specify custom dependency structures for simplex data. Moreover, the sum-to-one constraint induces negative correlations between components, making the Dirichlet unsuitable for scenarios where components may exhibit positive correlations (Connor and Mosimann, 1969; Wong, 1998). Even many complex negative correlation patterns cannot be accurately captured, further highlighting the limitations of the Dirichlet distribution in accommodating diverse dependency structures.

From a technical perspective, the Dirichlet distribution imposes additional constraints that limit its applicability. It requires properties such as closure under marginalization and conditioning, as well as complete subcompositional independence. This last property implies that all renormalized subsets of $\quad$ are independent of each other (Aitchison, 1986). Subcompositional independence has been widely criticized, as it imposes $1/2K(K-3)$ constraints on the covariance structure, rendering the Dirichlet implausible for most real-world applications (Aitchison and Shen, 1980; Aitchison, 1986; Ongaro and Migliorati, 2013). These limitations make the Dirichlet unsuitable for modeling complex or realistic dependency structures, necessitating more flexible alternatives like the Logistic Normal distribution, which we present below.

The Logistic Normal Distribution

The logistic normal distribution, proposed by Aitchison and Shen (1980) offers a solution to overcoming limited correlation structures. The core concept involves mapping a multivariate normal random variable defined from $\quad^K$ into the $K-1$ dimensional simplex $\quad^K$ by the use of log ratio transformations, as described below. This approach enables researchers to make use of well known established techniques for multivariate analysis in the unbounded real space and then seamlessly transforming the results into the simplex (Aitchison, 1986; Pawlowsky-Glahn and Buccianti, 2011).

Let $\quad \in \quad^K$, the *additive log-ratio* (alr) transformation, relative to the Kth component, is defined as: $\eta = \text{alr}(\quad) \coloneqq \left(\ln\left(\frac{\quad_1}{\quad_K}\right), \ldots, \ln\left(\frac{\quad_{K-1}}{\quad_K}\right)\right) \in \quad^{K-1}$ (Aitchison and

Shen, 1980). This transformation establishes a one-to-one correspondence between the elements of the simplex and the log-ratio vectors $\ln\left(\frac{i}{K}\right)$, allowing any statement about the components of the simplex to be equivalently expressed in terms of these log-ratios (Aitchison and Shen, 1980; Aitchison, 1986). By mapping compositional data into the unconstrained real space $^{K-1}$, the difficulties associated with working in a constrained space are effectively removed.

The alr transformation is inherently asymmetric since the log-ratios require a reference component. The last component $_K$ is usually chosen as reference, but any component can be selected based on interpretability or practical considerations (Aitchison, 1986). Guidelines for choosing the reference component have been proposed (Greenacre et al., 2023), but in general, practitioners of compositional data analysis (CDA) and correlated topic models (CTM) recognize that the choice of reference has minimal impact on results or inference (Pawlowsky-Glahn and Buccianti, 2011; Boogaart and Tolosana-Delgado, 2013; Greenacre et al., 2023).

The inverse alr transformation is defined by the *standard logistic* (*softmax*) transformation with a fill-up term for the reference component: $_k = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, k = 1, \ldots, K-1,$, $_K = 1 - \sum_{k=1}^{K-1} \ _k$. We say that $\in \ ^K$ follows an *additive logistic normal* (ALN) distribution if $\eta = \text{alr}(\ )$ follows a multivariate normal distribution in $^{K-1}$. The density of the ALN distribution is given by (Aitchison, 1986):

$$p(\ |\mu, \Sigma)| = |2\pi\Sigma|^{-1/2} \left(\prod_{k=1}^{K} \ _k \right)^{-1} \exp\left\{ -\frac{1}{2} \left[\text{alr}(\ ) - \mu\right]' \Sigma^{-1} \left[\text{alr}(\ ) - \mu\right] \right\}. \tag{9}$$

We denote this as $\sim \text{ALR}(\mu, \Sigma)$, where μ and Σ depend on the chosen reference component.

A symmetric representation for , independent of a reference component, can be obtained by starting from the unconstrained space and mapping into the simplex. Specifically, let $\eta \sim \ _K(\mu, \Sigma)$ and apply the *symmetric logistic transformation*:

$$_k = \frac{e^{\eta_k}}{\sum_{j=1}^{K} e^{\eta_j}}, \ k = 1, \ldots, K. \tag{10}$$

In this case, is said to follow a (symmetric) *Logistic Normal* (LN) distribution with parameters μ, Σ, denoted as $\text{LogisticNormal}(\mu, \Sigma)$. For identifiability, either the constraint $\sum_{k=1}^{K} \eta_k = 0$ or $\eta_K = 0$ can be imposed, the latter being equivalent to the ALR (Bishop, 2006; Blei and Lafferty, 2007; Goodfellow et al., 2016). If we consider the sum-to-zero constraint, then $\eta_k = \ln\left(\frac{k}{g(\ )}\right)$, where $g(\ )$ denotes the geometric mean of .

Importantly, $\Sigma = (\sigma_{ij}), i, j = 1, \ldots, K$, represents the covariance structure of the log-ratios, not of the raw proportions $\text{cov}(\ _i, \ _j)$. If one chooses to model using either the ALR or LN distribution, transitioning between them is entirely feasible via a straightforward linear relationships that we show in the Supplementary Material (Aguilar and

Bürkner, 2025). This flexibility allows researchers to select the representation most suited to the practical requirements of their analysis.

While moments of all non-negative orders, $\left(\prod_{k=1}^{K} \binom{m_k}{k}\right), m_k \geq 0, k = 1, \ldots, K,$ exist, their expressions are not reducible to simple analytic forms (Aitchison and Shen, 1980; Frederic and Lad, 2008; Holmes and Schofield, 2022). Nonetheless, when analyzing quantities in the simplex, it is customary to study ratios or logarithms of ratios of simplex components rather than raw proportions (Aitchison, 1986; Boogaart and Tolosana-Delgado, 2013; Pawlowsky-Glahn and Buccianti, 2011; Creus-Martí et al., 2021). The primary motivation for this is the so-called *coherence* requirement in compositional data analysis, which ensures that the relationships between components remain consistent and interpretable, whether considering the full composition or its subcompositions Greenacre et al. (2023). Coherence is especially important since the constant-sum constraint in compositional data implies that changes in one component inherently affect all others.

Let $\sigma_{ik,jl} = \mathrm{cov}\left(\ln\left(\ _i/\ _k\right), \ln\left(\ _j/\ _l\right)\right)$ represent the covariance of log-ratio quantities, then the following key relationships hold:

$$\left\{\ln\left(\ _j/\ _k\right)\right\} = \mu_j - \mu_k, \quad \sigma_{ik,jl} = \sigma_{ij} + \sigma_{kl} - \sigma_{il} - \sigma_{jk}. \tag{11}$$

Thus, the mean vector μ and covariance matrix Σ fully determine the expectations and pairwise covariances of log-ratio quantities. This implies that no recalculations are necessary when transitioning between different log-ratio representations, as all relevant information is encoded in μ and Σ. Moreover, one can derive μ and Σ directly from knowledge of any pair of log-ratio quantities, and vice versa (Aitchison, 1986). We explore these parameters in greater detail in Section 2.5, where we incorporate the LN distribution as a prior for within the GDR2 framework.

2.5 Hyperparameter Specification in GDR2 Priors

As explained in Section 2.3, the Beta prior for R^2 can be parametrized in terms of the prior mean μ_{R^2} and precision $_{R^2}$. In cases where the user has limited prior knowledge, a flat distribution over the unit interval can be specified by setting $(\mu_{R^2}, \ _{R^2}) = (0.5, 2)$. Alternatively, a bathtub-shaped distribution can be chosen with $(\mu_{R^2}, \ _{R^2}) = (0.5, 1)$, which places more mass on extreme values of R^2. This reflects a prior belief that either there is no relationship between the response and the covariates, or that the covariates fully explain the variability in the response. An application of the law of Total Variance shows that $\mathrm{var}(b) = \ [\mathrm{var}(b \mid \ , \omega^2)] + \mathrm{var}[\ (b| \ , \omega^2)] = \ (\omega^2) \ (\mathrm{cov}(\ ))$. Thus, the hyperparameters selected for ω^2 can control the tail behavior of b_k, potentially inducing infinite variance and heavy tails in its marginal distribution. This configuration enhances sensitivity to detecting signals.

As discussed in Section 2.4, we let follow either a Dirichlet or an LN distribution. When setting $\sim \mathrm{Dirichlet}(\alpha)$ the GDR2 prior is equivalent to the R2D2 prior (Zhang et al., 2020) for non-hierarchical models, except that b_k follows a normal distribution rather than a Double Exponential. The former scenario has been explored by Aguilar

and Bürkner (2023), who show that if $\sim \text{Dirichlet}(\alpha)$ with $\alpha = (a_\pi, \ldots, a_\pi)$, then a_π determines the behavior of the marginal distribution of b_k near the origin. Thus, a_π can be theoretically chosen to tailor the properties of the marginal distribution of b_k. As an alternative, the elements of α can be elicited from subject matter experts to represent the a priori relative importance of covariates – although we consider it very hard elicit such knowledge for high-dimensional problems.

If $\sim \text{LogisticNormal}(\mu, \Sigma)$, the prior specification involves defining the mean vector μ and covariance matrix Σ for the log-ratios. Setting $\mu_k = 0\ \forall k$ (Equation (11)) encodes the belief that all proportions $_k$ are equally weighted, expressing no preference for any specific component. To assign equal weights to a subset of proportions, $\mu_i = c$ can be specified for $i \in I \subset \{1, \ldots, K-1\}$. If the proportions can be grouped into G mutually exclusive sets $I_g \subset \{1, \ldots, K-1\}$, group-specific weights c_g may reflect their relative importance, enabling the inclusion of prior hierarchical or structural information.

To better understand the effect of assigning values to μ, consider the symmetric logistic mapping in Equation (10) and assume $\Sigma = \sigma^2 I$. Setting $\mu_k = c\ \forall k$ is equivalent to $\mu_k = 0$, as the logistic mapping is shift-invariant (Bishop, 2006). This choice reflects a lack of prior knowledge about relative importance and by symmetry it can be shown that $[\ _k] = 1/K$. When $\mu_k = c_k$, proportions will differ depending on the sign of c_k; components with $c_k > 0$ will, on average, be larger, while those with $c_k < 0$ smaller. To see this, let $R_k = e^{\eta_k}$ and $S = \sum_j e^{\eta_j}$. Using a first order Taylor expansion around the mean, we can approximate (Stuart and Ord, 2009):

$$[\ _k] = \left(\frac{e^{\eta_k}}{\sum_j e^{\eta_j}}\right) \approx \frac{[R_k]}{[S]} = \frac{e^{c_k + \frac{\sigma^2}{2}}}{\sum_j e^{c_j + \frac{\sigma^2}{2}}} = \frac{e^{c_k}}{\sum_j e^{c_j}}, \tag{12}$$

where we compute the expectations by noting that R_k follows a log normal distribution. This argument extends to a diagonal covariance matrix Σ with different scales σ_j; whose values interact with c_j to redistribute mass in favor or against $_k$.

The relationship becomes more complex for a general Σ and μ. While a Taylor expansion provides an approximation,[1] it relies on the term $\text{Cov}(R_k, S)$, which lacks a closed-form solution, requiring Monte Carlo methods for estimation. While this offers insights into the interaction of key quantities that determine the marginal expectation, its added complexity may outweigh its utility. Our experiments in Section 3 show that, setting $\mu_k = 0$ is a robust default choice in the absence of prior knowledge, balancing simplicity and performance. The specification of Σ can be understood by analyzing the implied prior on the norm of η (the log ratios) and how this translates to the simplex. We first consider the case of a diagonal $\mu = 0, \Sigma = \sigma^2 I$. Let $\eta \sim N(0, \sigma^2 I)$, since η is a sub-gaussian variable with independent entries (Giraud, 2014; Vershynin, 2018), $\|\eta\|$ concentrates around $\sqrt{K}\sigma$. Specifically, the following two-sided tail bound holds: $\left(\left|\frac{\|\eta\|}{\sqrt{K}\sigma} - 1\right| \geq t\right) \leq 2\exp\left(-\frac{K\sigma^2 t^2}{2C}\right)$ for a constant $C > 0$. We illustrate this behavior in Figure 4.

[1] see Supplementary Material (Aguilar and Bürkner, 2025).

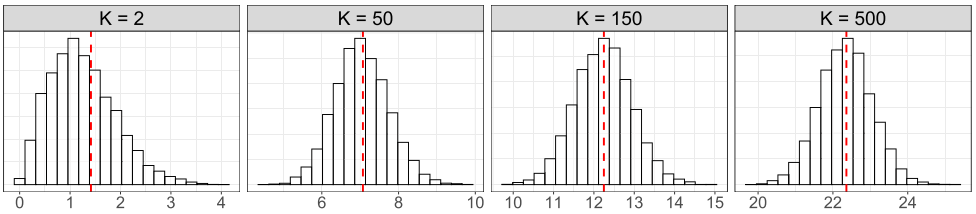


Figure 4: Implied distribution on the norm of a Multivariate Normal as the dimension K increases. The red line shows the value $\sqrt{K}\sigma$, which is the point of concentration, where K is the number of covariates and $\sigma = 1$.

The user-specified σ defines bounds for the norm of the logits, setting a "budget" that can be distributed across them. For $\sigma \to 0$, the logits concentrate near zero, resulting in $_k \to 1/K$. Conversely, as σ increases, the logits spread further, increasing the quota for $\|\eta\|$. If $\mu \neq 0$, nonsparse elements will be determined by the locations in which $\mu_k \neq 0$, especially if the budget is low.

The choice of σ balances the variability in the simplex as K grows. Setting $\sigma = \sqrt{\gamma/K}, \gamma > 0$ ensures that $\|\eta\|^2 \approx \gamma$, maintaining a consistent magnitude for the logits regardless of K. Alternatively, choosing $\sigma = \sqrt{\gamma}$ (independent of K) allows the logits' norm to grow proportionally to $\sqrt{K}$, introducing greater variability. This variability interacts with μ to influence how mass is distributed across $_k$. For scenarios where sparsity is critical, setting $\sigma = \sqrt{\log(K)/K}$ slows the growth of the logits' norm, preventing excessively large values and reducing variability. This encourages sparse distributions, with mass concentrated at locations dictated by μ. Such a setup is particularly effective in high-dimensional settings. Users must carefully consider the interaction between μ and σ . Lower values of σ concentrate the mass of at positions where μ has higher values, promoting focused distributions. In contrast, higher values of σ introduce variability, allowing for more diffuse distributions across the simplex.

When Σ incorporates non-zero correlations, deriving a general inequality for the concentration of measure of $\|\eta\|$ becomes significantly more complex. We do not explore this case further, since we consider the intuition provided by σ sufficient for practical specification. However, if users have prior knowledge about how the variables are correlated, they can include this information via a correlation matrix Ω , using the decomposition $\Sigma = \mathrm{D}\ \Omega\ \mathrm{D}$, where D is a diagonal matrix containing the scales, and Ω is a correlation matrix. For example, if the coefficients are ordered and proportions are expected to be similar for adjacent variables, one might use an autoregressive correlation structure. Alternatively, for cases where groups of variables exhibit distinct interdependencies, a block correlation matrix may be appropriate. This flexibility allows users to encode meaningful structural relationships directly into the prior.

Prior Matching

An alternative approach to specifying the hyperparameters of the LN distribution is to draw upon theoretical insights and practical experience from the $\sim \text{Dirichlet}(\alpha)$ case. Specifically, we propose using the Dirichlet distribution as an initial guide to determine the hyperparameters of the LN distribution. Our method involves defining a Dirichlet distribution that captures the desired theoretical and practical characteristics of the prior. Then, we minimize a divergence measure between the Dirichlet and LN distributions. This automated process, which we term *prior matching*, provides a systematic way to derive the prior mean and covariance matrix for the LN distribution.

Prior matching simplifies the complex task of specifying LN hyperparameters by using the more analytically tractable Dirichlet distribution as a reference. While the LN distribution offers increased flexibility compared to the Dirichlet, its analytical properties can be challenging to study, particularly in the context of shrinkage priors. Prior matching mitigates these challenges, automating the derivation of the LN's prior mean and covariance matrix for improved efficiency and user convenience. Exact matching is not the ultimate goal, as our objective is to develop priors that outperform the Dirichlet. Prior matching serves as a starting point, enabling users to leverage the unique flexibility of the LN distribution. For instance, the LN distribution addresses a critical limitation of the symmetric Dirichlet distribution: sampling difficulties and numerical instability as $a_\pi \to 0$.

As divergence metric for prior matching, we propose to use the Kullback Leibler (KL) divergence (Kullback and Leibler, 1951). The KL divergence between two probability distributions f, g is defined as $\text{KL}(f||g) := {}_f\left[\ln\left(\frac{f(x)}{g(x)}\right)\right] = \int_{-\infty}^{\infty} \ln\left(\frac{f(x)}{g(x)}\right) f(x)dx$. To find the LN distribution that best approximates a given Dirichlet distribution, we minimize the KL divergence between them. Letting $f \sim \text{Dirichlet}(\alpha)$ with α fixed and $g \sim \text{LogisticNormal}(\mu, \Sigma)$, the optimal LN parameters are obtained by solving: $\mu^*, \Sigma^* = \arg\min_{\mu,\Sigma} \text{KL}(f||g)$. The closed form solution is given by

$$\mu_k^* = \delta(\alpha_k) - \delta(\alpha_K), \sigma_{kk}^* = \varepsilon(\alpha_k) + \varepsilon(\alpha_K),\ k = 1, \ldots, K,\quad \sigma_{kj}^* = \varepsilon(\alpha_K)\ (k \neq j), \quad (13)$$

where $\delta(x) = \frac{\Gamma'(x)}{\Gamma(x)}$ and $\varepsilon(x) = \delta'(x)$ are the digamma and trigamma functions respectively (Aitchison and Shen, 1980; Aitchison, 1986; Jeffrey et al., 2007). Given their analytical form, these quantities can be computed instantly, adding no computational overhead to the procedure. While other divergence measures could be considered, the closed-form solution for the KL divergence offers a clear advantage in terms of both practical applicability and computational efficiency.

To exemplify KL matching, consider a symmetric Dirichlet distribution with a single hyperparameter a_π. The value of a_π controls the amount of shrinkage induced by the R2D2 prior (Aguilar and Bürkner, 2023). When $a_\pi \leq 1/2$ the prior will be unbounded near the origin, leading to strong shrinkage towards zero. Users who wish to replicate this shrinkage behavior in the LN distribution can propose a value of $a_\pi \leq 1/2$ and use Equation (13) to calculate the corresponding LN hyperparameters: $\mu_k = 0, \sigma_{kk} =$

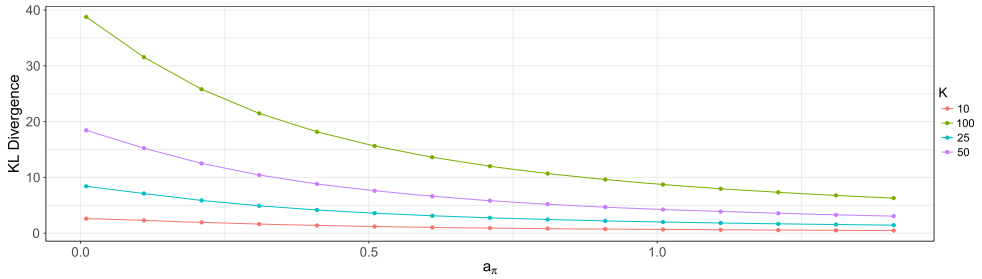


Figure 5: KL divergence between a symmetric Dirichlet distribution (parameter a_π) and the closest logistic normal distribution as a_π and the number of covariates (K) vary. KL divergence decreases monotonically with a_π.

$2\varepsilon(a_\pi)$, and $\sigma_{kj} = \varepsilon(a_\pi)$. As a_π increases, the KL divergence between the Dirichlet and LN distributions decreases (Aitchison, 1986). When $a_\pi \to 0$, the KL divergence increases significantly since the Dirichlet allocates mass along the simplex edges, a challenge for the LN distribution to approximate, given that it is defined within the simplex's interior. This is illustrated in Figure 5 for varying values of a_π and different dimensions K.

A straightforward modification to KL matching involves discarding the correlations in the derived Σ^*, using only its diagonal elements to create a diagonal covariance matrix $\widetilde{\Sigma}^*$. In this case the scales $\sigma^2_{kk} = 2\varepsilon(a_\pi), \forall k$ and $\text{cor}(\eta_i, \eta_j) = 1/2, i \neq j$. In high dimensional settings, it is unrealistic to expect such correlations as a default, as we typically anticipate only a small number of signals (Tosh et al., 2022). Extracting only the scales, while discarding the correlations, is sensible since this type of correlations may be rendered as noise or irrealistic in high dimensions. The deviation between $\widetilde{\Sigma}^*$ and Σ^* can be quantified by the KL divergence in log-ratio space. In the case where Σ^* is obtained through KL matching with a symmetric Dirichlet distribution with hyperparameter a_π, the value of divergence has a simple form: $\frac{1}{2} \ln \left(\frac{2^{K-1}}{K} \right)$, which is independent of a_π.

We conclude this section by noting that discussions on selecting hyperparameter values for the Logistic Normal distribution as a prior are limited, and practical guidance is often scarce. Researchers often prefer to set priors over the hyperparameters (Blei and Lafferty, 2007; Mimno et al., 2008; Xun et al., 2017), either by using priors of the form $\mu \mid \Sigma \sim \quad (\mu_0, \gamma\Sigma)$ with $\Sigma \mid W \sim \text{InverseWishart}(\delta, W^{-1})$ (Chen et al., 2013), or by employing a Variational Approximation to the posterior distribution, where the variational family assumes a diagonal structure for Σ (Blei and Lafferty, 2007; Blei et al., 2017). However, these methods often lack sufficient discussion on the choice of hyperpriors for the hyperparameters. In this work, we opted not to explore these approaches due to their added computational complexity, comparable performance with fixed hyperparameters, and the limited understanding of how such specifications interact with shrinkage priors in fixed-sample settings.

3 Experiments and Case Studies

We conducted simulation studies to assess the performance of our generalized R2 priors under various conditions commonly encountered in practical scenarios. These investigations provide insights into how the priors behaves with finite samples and explores different possibilities for the total variance decomposition. Additionally, we present case studies utilizing real-world datasets that are commonly used as benchmarks for shrinkage priors. The shrinkage priors considered in our simulations include the Beta Prime, Dirichlet-Laplace, Horseshoe, and the proposed GDR2 model, which incorporates Dirichlet and Logistic Normal decompositions. While these priors provide a useful context for comparison, our primary focus is on the GDR2 model, particularly its variance decomposition mechanisms. Results for the other priors, evaluated using their default settings, are detailed in Aguilar and Bürkner (2025). Notably, the GDR2 model demonstrates on par performance to these well-established priors and even outperforms them in specific scenarios. This highlights the robustness of GDR2 and underscores the flexibility and potential of decomposition-based approaches.

All models were implemented in the probabilistic programming language Stan (Carpenter et al., 2017; Stan Development Team, 2024), which provides an extended implementation of the No-U-Turn Sampler (NUTS) (Hoffman and Gelman, 2011), an adaptive form of Hamiltonian Monte Carlo (HMC) (Brooks et al., 2011). The Stan code for the GDR2 prior is included in the Supplementary Material (Aguilar and Bürkner, 2025), and all associated data and code can be found in https://osf.io/ns2cv/.

3.1 Simulations

Generative Models

We use model (1) as a generative model in our simulations, adapting it to accommodate various forms of data encountered in practice, including diverse levels of sparsity and different dependency structures among the covariates. The design matrix X was sampled from a multivariate normal distribution with mean 0 and covariance matrix Σ_x derived from an AR(1) correlation structure with $\rho_x \in \{0, 0.8\}$. The intercept b_0 was drawn from a normal distribution with mean zero and variance $\sigma_I^2 = 4$. We fixed the sample size at $N = 100$, and varied the number of regression coefficients $K \in \{50, 150, 750\}$ to represent both low-dimensional ($K < N$) and high-dimensional ($K \gg N$) scenarios. The residual standard deviation σ was adjusted using Equation (4) to maintain the true explained variance $R_0^2 \in \{0.25, 0.6\}$.

The regression coefficients b_k, were generated in two ways: 1) *Simulated Coefficients:* Coefficients were sampled from a normal distribution with zero mean and covariance Σ_b. Two forms of Σ_b were considered: (A) diagonal with $\sigma_b^2 = 9$, and (B) AR(1) with autocorrelation $\rho_b = 0.8$ and diagonal elements $\sigma_b^2 = 9$. Sparsity was induced by setting each coefficient to zero with probability $v = 0.75$, reflecting real-world scenarios where coefficients arise from random processes (Bhattacharya et al., 2015; Griffin and Brown, 2010, 2013). 2) *Fixed Coefficients*: This setup is akin to examples encountered in the shrinkage prior literature (Carvalho et al., 2010; Griffin and Brown, 2010; Bhattacharya

et al., 2015; Zhang et al., 2020). We place concentrated signals $b^* \in \{3, 7\}$ in specific locations of the coefficient vector b. The first 5 and the last 5 elements are set to b^*, while all others are set to 0.

We have specified Dirichlet and symmetric Logistic Normal distributions for the proportions of total variance . Two hyperparameter configurations for R^2 were tested: *default* $(\mu_{R^2}, \ _{R^2}) = (0.5, 1)$ and *uniform* $(0.5, 2)$. When let follow a symmetric Dirichlet distribution with concentration parameter $a_\pi \in \{0.5, 1\}$. The *default* case $a_\pi = 0.5$ encourages shrinkage by pushing mass toward the edges of the simplex, whereas the "*uniform*" case $a_\pi = 1$ reduces shrinkage. When $a_\pi \leq 0.5$, the marginal prior distributions of the coefficients b_k become unbounded near the origin, enforcing sparsity by shifting mass to the simplex edges (Aguilar and Bürkner, 2023).

To specify the hyperparameters of $\sim \text{LogisticNormal}(\mu, \Sigma)$, we used KL matching (Section 2.5), referring to the outcome as *Full KL*. Specifically, for $a_\pi = 0.5$, we set $\mu_k = 0$, $\sigma_{kk} = \pi$, and $\sigma_{kj} = \frac{\sqrt{2}}{2}\pi$; for $a_\pi = 1$, we set $\mu_k = 0$, $\sigma_{kk} = \frac{\sqrt{3}}{3}\pi$, and $\sigma_{kj} = \frac{\sqrt{6}}{6}\pi$, both yielding pairwise correlations of 0.5. We also consider a variant restricting Σ^* to a diagonal matrix, termed *Scales KL*. Table 1 summarizes the decompositions.

As discussed in Section 2.5, the marginal distributions of for the Logistic Normal (LN) priors may exhibit spikes around $1/K$ as K increases and $\mu = 0$. When using diagonal covariance matrices, the LN prior allows for greater variability in the marginals, providing more flexibility, serving as a less informative prior in comparison. If prior knowledge about the location of strong signals exists, this should be encoded in μ. In this study, we mimic complex scenarios with minimal prior knowledge about the coefficients to test the prior under such conditions.

We crossed the default and uniform hyperparameter specifications for R^2 and a_π, resulting in four hyperparameter specifications. We present a subset of the results for the default-default case $(\mu_{R^2}, \ _{R^2}, a_\pi) = (0.5, 1, 0.5)$ and show others in the Supplementary Material (Aguilar and Bürkner, 2025). Additionally, we include results for an informative case in the fixed coefficient setup in Supplementary Material (Aguilar and Bürkner, 2025), where α is constructed in a way that incorporates strong user knowledge to emphasize important signals, matching the locations of the signals. This helps assess the extent to which both decompositions benefit from user-provided information.

After fully crossing all conditions, we obtained a total of 192 different simulation configurations. For each configuration, we generated $T = 100$ datasets consisting of $N = 100$ training observations y_n, $n = 1, \ldots, N$. Predictive metrics (discussed below) were computed based on $N_{\text{test}} = 500$ test observations, which were generated independently

Name	Label	Distribution	Details
Dirichlet	D2	Dirichlet	$\alpha_k = a_\pi$
Full KL	LNF	logistic normal	μ^*, Σ^* from KL matching
Scales KL	LNS	logistic normal	$\mu^*, \text{diag}(\Sigma^*)$ from KL matching

Table 1: Different decompositions that are considered in the simulations.

from the training data using the same data-generating mechanism. The simulations were carried out on the Linux HPC Cluster (LiDO3) at TU Dortmund University.

Evaluation Metrics

We evaluated and compared the performance of the models based on two criteria: out-of-sample predictive performance and parameter recovery (Vehtari and Ojanen, 2012; Robert, 2007). Out-of-sample predictive performance was assessed using the expected log-pointwise predictive density (ELPD) (Vehtari and Ojanen, 2012), computed on the test data as $\text{ELPD} = \sum_{i=1}^{N} \ln \left(\frac{1}{S} \sum_{s=1}^{S} p(y_i|\theta^{(s)}) \right)$, where $\theta^{(s)}$ represents the sth posterior draw for $s = 1, \ldots, S$. The ELPD quantifies predictive accuracy across the N data points, evaluating the quality of the overall predictive distribution. A higher ELPD value indicates better out-of-sample predictive performance. Log probability scores, such as ELPD, are often recommended as a general-purpose metric when there is no specific reason to use an alternative (Vehtari and Ojanen, 2012; Vehtari et al., 2016, 2022).

Parameter recovery was evaluated using the posterior Root Mean-Squared Error (RMSE), calculated as $\text{RMSE} = \frac{1}{K} \sum_{k=1}^{K} \sqrt{\frac{1}{S} \sum_{s=1}^{S} \left(b_k^{(s)} - b_k \right)^2}$, where $b_k^{(s)}$ represents the sth posterior draw for $k = 1, \ldots, K$, and b_k denotes the true value of the kth regression coefficient. The RMSE serves as a comprehensive measure of estimation error, naturally capturing the tradeoff between bias and variance. In our analysis, we compute three versions of RMSE: (1) Averaged over all coefficients, (2) only over the truly zero coefficients, (3) only over the truly nonzero coefficients. To provide a more comprehensive evaluation of the posterior inference properties, we assess the coverage of 95% marginal credible intervals for each approach based on average interval width, coverage proportion, specificity, and sensitivity. We also present Receiver Operating Characteristic (ROC) curves. These results are presented in the Supplementary Material (Aguilar and Bürkner, 2025).

To compare multiple models based on a metric of interest (e.g., ELPD, RMSE), we computed the difference relative to the best-performing model for a given dataset. Let $\{m_1, \ldots, m_l\}$ represent the set of models, and denote the best model with respect to by m^*. The Delta metric, Δ_i, for the ith model is defined as $\Delta_i = (m_i) - (m^*)$, where $i = 1, \ldots, l$. This approach highlights differences between models while accounting for variations caused by randomness in the simulated datasets, thus focusing the evaluation on meaningful distinctions.

Metrics for diagnosing and evaluating the Markov Chain Monte Carlo (MCMC) sampler, including Effective Sample Size (ESS) and Rhat (Geyer, 1992; Brooks et al., 2011; Vehtari et al., 2021), are provided in the Supplementary Material (Aguilar and Bürkner, 2025). Overall, the results indicate effective sampling from the posterior, as ESS values are sufficiently high and Rhat values fall within acceptable ranges. Additionally, all models demonstrate comparable computational efficiency, with similar runtimes observed across the board.

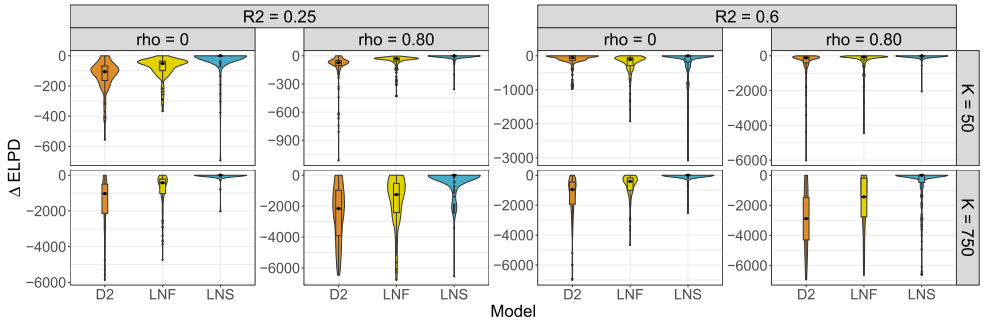


Figure 6: **Simulated coefficients setup.** ΔELPD evaluated on test datasets of size $N = 500$.

Results: Simulated Coefficients Setup

We present results for the scenario where the coefficients are generated with an AR(1) correlation structure, characterized by $\rho_b = 0.8$ and marginal variances $\sigma_b^2 = 9$, alongside an average sparsity of $\nu = 0.75$. In real-world applications, it is reasonable to encounter clusters of coefficients that are correlated with each other, even though these correlations do not necessarily translate into significant effects on the response variable. Figure 6 shows the distribution of ΔELPD across simulations under various conditions, visualized as violin plots with embedded boxplots (Hintze and Nelson, 1998; Stryjewski, 2010). The results demonstrate that Logistic Normal (LN) decompositions consistently outperform their Dirichlet counterparts. Notably, the LN priors exhibit the best average performance across all scenarios. The LNS version is the clear winner in predictive performance.

Figure 7 depicts the distribution of ΔRMSE across various simulation conditions. Similar to predictive performance, the LN priors demonstrate the most robust parameter recovery among the competing R^2-based models, regardless of the values of ρ, R^2, or K. The overall RMSE results reflect a balance between the model's posterior error for truly zero and nonzero coefficients. To explore this further, Figure 7 separately illustrates ΔRMSE for truly zero and nonzero coefficients.

These results suggest that LN decompositions achieve more effective shrinkage compared to Dirichlet counterparts, reducing false positives and improving the detection of truly zero coefficients. Dirichlet decompositions impose less shrinkage, which can result in poorer performance when coefficients are truly zero. The trends for nonzero coefficients are similar. Of particular note is the case with $K = 750$, a challenging scenario where shrinkage priors are expected to overshrink. In this high-dimensional setting, LN decompositions show improved behavior when moving from $\rho = 0$ to $\rho = 0.80$. This improvement is also present when transitioning from a low signal ($R^2 = 0.25$) to a moderate signal ($R^2 = 0.60$). Overall, LN decompositions offer a more robust and adaptable approach to shrinkage in sparse, high-dimensional contexts.

In summary, using a zero mean vector and a diagonal covariance matrix for LN

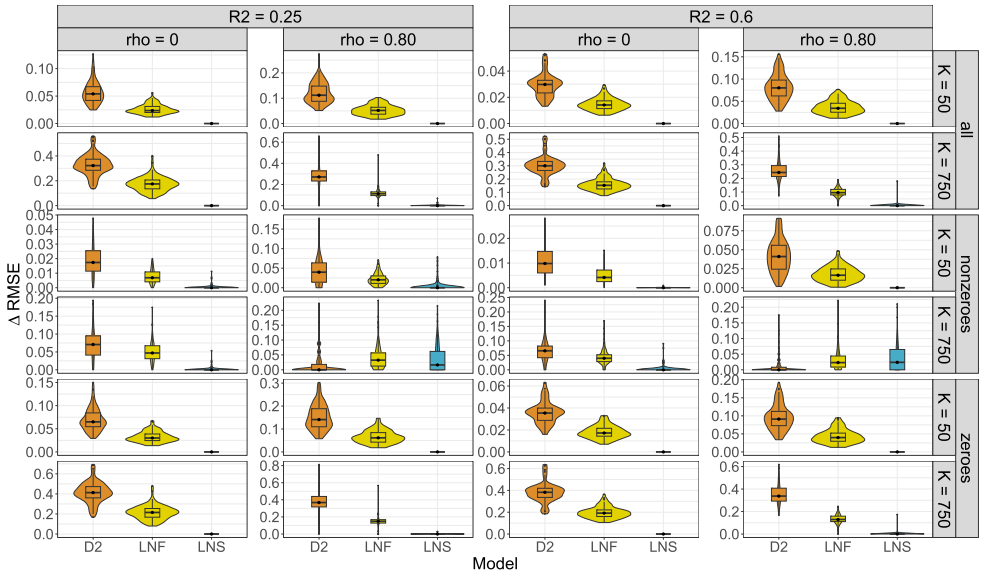


Figure 7: **Simulated coefficients setup.** Violin plots with embedded box plots for ΔRMSE under different simulation conditions. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

priors appears to be an effective strategy, particularly when prior knowledge about relationships between coefficients is limited—a common scenario in practice. These results highlight the potential of LN-based approaches as robust default alternatives to Dirichlet-based shrinkage priors. Even with default parameter settings, LN priors demonstrate competitive or superior performance across diverse scenarios.

Results: Fixed Coefficients Setup

We show the results when the fixed value is equal to 3. Figure 8 reveals that the LN decompositions consistently excel in out-of-sample predictive performance when compared to the Dirichlet counterpart. The LNS stands out as the most efficient predictive model. The most pronounced disparities emerge in scenarios characterized by high correlation and dimensionality ($\rho = 0.8, K = 750$), where the performance of the Dirichlet prior is notably subpar. The results for ΔRMSE for all of the coefficients and the zero coefficients behave similarly to the simulated coefficients setup. The LN decompositions perform clearly and uniformly better the Dirichlet for all the coefficients and for the zero coefficient only. This indicates again that LN decompositions are better at both overall RMSE and noise detection. We only show the case for nonzero coefficients in Figure 9. A complete overview over the other cases is available in the Supplementary Material (Aguilar and Bürkner, 2025).

The Dirichlet decomposition excels only in the in low-signal, high-correlation set-

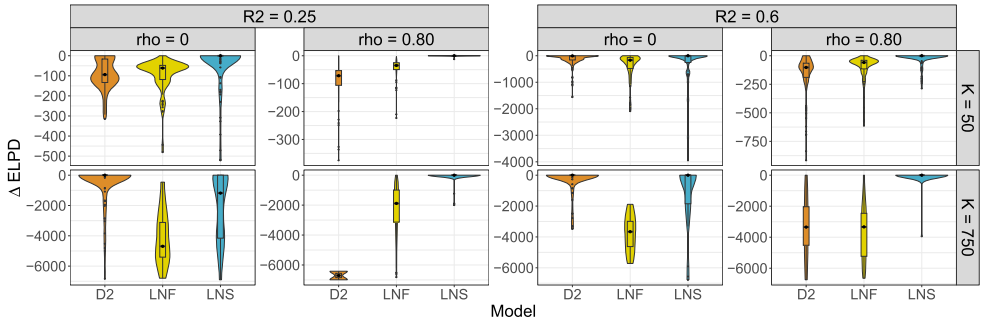


Figure 8: **Fixed coefficients setup.** ΔELPD evaluated on the test datasets for the fixed coefficients when the signal have a value of 3.

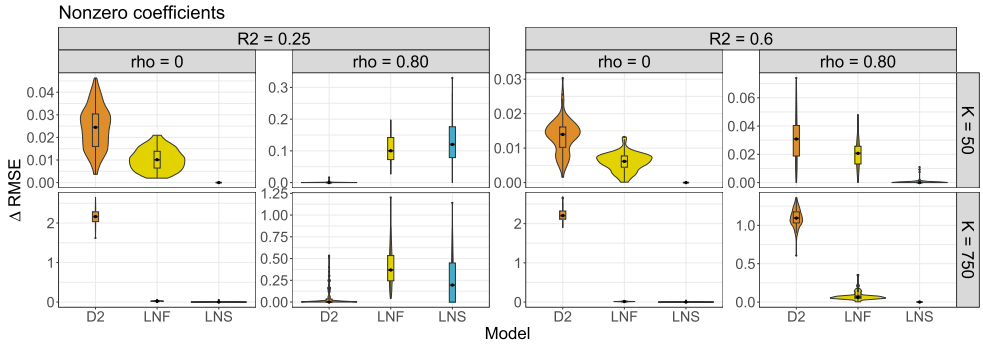


Figure 9: **Fixed coefficients setup.** ΔRMSE for truly nonzero coefficients.

tings. LN decompositions achieves better performance in all other scenarios. These results suggest that the effectiveness of LN decompositions could be enhanced even more by incorporating prior structure that mitigates correlations in the design matrix. If we focus on the $\rho = 0.8, K = 750$ case in both Figures 8 and 9, we can see that there is a tradeoff between out-of-sample predictive performance and detecting signals in that case. The Dirichlet lacks sufficient shrinkage (worsening out-of-sample predictions), while the LN may overshrink some signals but improves them overall.

3.2 Real-World Case Studies

We assess the predictive performance of our proposed prior using three high-dimensional real-world datasets, each with distinct correlation structures among covariates. The Cereal dataset contains starch content measurements for 15 observations with 145 infrared spectra predictors, sourced from the R package `chemometrics` (Filzmoser, 2023). The Cookie dataset includes fat content measurements from 72 samples with 700 Near - infrared spectra predictors, originally introduced by Osborne et al. (1984) and available

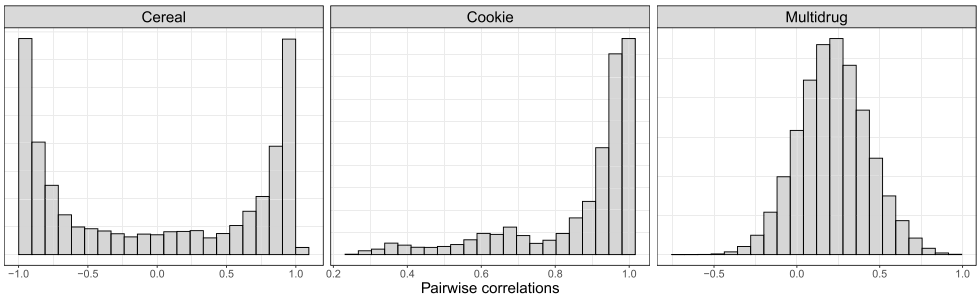


Figure 10: Histograms of pairwise correlations of the covariates of the different datasets being considered.

in the orphaned `ppls` R package (Boulesteix, 2014). The Multidrug dataset, derived from a pharmacogenomic study on drug concentration and gene expression (Szakács et al., 2004), comprises 60 observations and 853 covariates after removing missing values. The Cereal and Multidrug datasets have been analyzed in prior work on shrinkage priors (Polson and Scott, 2011; Griffin and Brown, 2013; Zhang et al., 2020), while the Cookie dataset has been studied in Osborne et al. (1984); Ghosh and Ghattas (2015); Zhang et al. (2020). These datasets exhibit varied predictor dependencies: mixed correlations in Cereal, strong positive correlations in Cookie, and low-to-moderate correlations in Multidrug (Zhang et al., 2020), as visualized in Figure 10.

We use ELPD to quantify out-of-sample predictive performance. Since we don't have an independent test set available, we use cross-validation to estimate out-of-sample ELPD as follows: We split each dataset into training and test sets, with 75% of observations used for training and the remaining 25% for testing. We repeat this 100 times and use the resulting ELPD values, which are subsequently summed to obtain an overall ELPD estimate. We compare the performance of the Dirichlet, LNF, and LNS decompositions with specific parameter settings, namely $(\mu_{R^2}, \quad _{R^2}, a_\pi) = (0.5, 1, 0.5)$.

We summarise the results in Table 2. LNS consistently demonstrates superior ELPD performance compared to both Dirichlet and LNF across the diverse datasets. The latter

Dataset	Cereal ($N = 15, K = 145$)	Cookie ($N = 72, K = 700$)	Multidrug ($N = 60, K = 853$)
Model	ΔELPD (SE)	ΔELPD (SE)	ΔELPD (SE)
LNS	0.0 (0.0)	0.0 (0.0)	0.0 (0.0)
LNF	−101.4 (85.1)	−673.1 (471.5)	−2950.4 (743.8)
D2	−96.8 (42.2)	−518.9 (274.4)	−5464.2 (794.7)

Table 2: Differences in ELPD and standard deviations for the different datasets considered, computed through pairwise comparisons with the model having the highest ELPD (in the first row). The initial value is zero, and subsequent rows display negative values that indicate the difference with the best model (Vehtari et al., 2023). See Vehtari et al. (2016) for details on standard error calculations.

two priors clearly differ from each other only for the Multidrug data where LNF was better than Dirichlet. This showcases that priors that exert more shrinkage on complex, high-dimensional datasets can be beneficial in terms of predictive performance.

4 Discussion

We introduced the Generalized Decomposition R2 prior framework for high-dimensional Bayesian regression, building on and extending the R2D2 prior (Zhang et al., 2020; Aguilar and Bürkner, 2023). Our key innovation lies in allowing to vary the prior used for the explained variance decomposition, enabling a more nuanced exploration of dependency structures. Previous methods are confined to Dirichlet decompositions, and thus can only express negative covariances tied to the means of the explained variance proportions. We overcome this limitation by employing the logistic normal distribution, facilitating the expression of covariance structures by transitioning from the simplex to the unconstrained real space.

Our simulations and real-world case studies demonstrate substantial gains in predictive performance when using logistic normal priors for the proportions of explained variance. The flexible dependency structures encoded in the logistic normal allow for a broader range of prior assumptions compared to the Dirichlet distribution. Specifically, we show that use of logistic normal based decompositions leads to superior out-of-sample predictive performance compared to the use of Dirichlet counterparts. In terms of parameter recovery, the Logistic normal prior models showed on par and superior performance compared to Dirichlet in most scenarios.

Our primary goal was to build well-predicting models, using the full set of available covariates and strong shrinkage on the corresponding coefficients. That said, as is the case for continuous shrinkage priors more generally, they do not create exact sparsity directly. Rather, redundant coefficients are shrunk to values close to zero, but not to zero exactly. As a result, predictive capabilities are not necessarily directly visible in few nonzero coefficients but rather implicitly distributed also among the almost zero coefficients, at least in high-dimensional settings (Piironen and Vehtari, 2017). In other words, we argue that our models (and other continuous shrinkage priors models) should not be used directly for variable selection (for more details and discussion see Piironen et al. (2020); Zhang and Bondell (2018); Tadesse and Vannucci (2021)). Rather, we recommend a two-step procedure where, after fitting a well predicting reference model with all covariates in the first step, we apply a separate, dedicated variable selection procedure in the second step, for example, projection predictive variable selection (Piironen et al., 2020; Catalina et al., 2020; McLatchie et al., 2023).

In future research, one important direction will be to study the influence and optimal choice of the logistic normal hyperparameters in more detail. While the hyperparameter choices proposed here already showed strongly improved performance compared to Dirichlet, it is still unclear how much room for improvement remains within the class of logistic normal decomposition priors. One promising direction we see in this context is the inclusion of covariate grouping or covariate dependency information in the prior

(Boss et al., 2023). But exactly how this information should be encoded, and if its inclusion is actually beneficial for the prior's performance, remains to be studied.

Acknowledgments

We would like to thank Aki Vehtari and his research group at Aalto University for their valuable comments and insightful discussions on an earlier version of this work. In particular, we are grateful to David Kohns for his detailed and constructive feedback.

Funding

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy EXC 2075-390740016 and DFG Project 500663361. We acknowledge the support by the Stuttgart Center for Simulation Science (SimTech). We acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by DFG Project 271512359.

Supplementary Material

Supplementary Material for: "Generalized Decomposition Priors on R2" (Aguilar and Bürkner, 2025) (DOI: 10.1214/25-BA1524SUPP; .pdf). In this supplementary material, we present further derivations and results.

References

Agresti, A. and Hitchcock, D. B. (2005). "Bayesian inference for categorical data analysis." *Statistical Methods and Applications*, 14(3): 297–330. MR2211337. doi: https://doi.org/10.1007/s10260-005-0121-y. 1494, 1502

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711–1767. MR4609453. doi: https://doi.org/10.1214/23-EJS2136. 1494, 1495, 1497, 1498, 1499, 1501, 1505, 1508, 1511, 1517

Aguilar, J. E. and Bürkner, P.-C. (2025). "Supplement to "Generalized Decomposition Priors on R2"." *Bayesian Analysis*. doi: https://doi.org/10.1214/25-BA1523SUPP. 1495, 1504, 1506, 1510, 1511, 1512, 1514, 1518

Aitchison, J. and Shen, S. M. (1980). "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika*, 67(2): 261–272. URL http://www.jstor.org/stable/2335470 MR0581723. doi: https://doi.org/10.2307/2335470. 1503, 1504, 1505, 1508

Aitchison, J. J. (1986). *The statistical analysis of compositional data / J. Aitchison.*. Monographs on statistics and applied probability (Series). Chapman and Hall. MR0865647. doi: https://doi.org/10.1007/978-94-009-4109-0. 1494, 1500, 1502, 1503, 1504, 1505, 1508, 1509

Armagan, A., Clyde, M., and Dunson, D. B. (2011). "Generalized beta mixtures of Gaussians." In *Advances in neural information processing systems*, 523–531. 1494

Armagan, A., Dunson, D. B., and Lee, J. (2013). "Generalized double Pareto shrinkage." *Statistica Sinica*, 23(1): 119. MR3076161. 1494, 1496

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233. MR4034859. doi: https://doi.org/10.1080/02331888.2019.1662017. 1497

Barndorff-Nielsen, O. and Jørgensen, B. (1991). "Some parametric models on the simplex." *Journal of Multivariate Analysis*, 39(1): 106–116. MR1128675. doi: https://doi.org/10.1016/0047-259X(91)90008-P. 1502

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2017). "The Horseshoe+ Estimator of Ultra-Sparse Signals." *Bayesian Analysis*, 12(4): 1105–1131. MR3724980. doi: https://doi.org/10.1214/16-BA1028. 1494, 1496, 1497

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2019). "Lasso Meets Horseshoe: A Survey." *Statistical Science*, 34(3): 405–427. MR4017521. doi: https://doi.org/10.1214/19-STS700. 1494, 1497

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. PMID: 27019543. MR3449048. doi: https://doi.org/10.1080/01621459.2014.960967. 1494, 1496, 1501, 1510

Bishop, C. (2006). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer. URL https://books.google.de/books?id=kTNoQgAACAAJ MR2247587. doi: https://doi.org/10.1007/978-0-387-45528-0. 1504, 1506

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). "Variational Inference: A Review for Statisticians." *Journal of the American Statistical Association*, 112(518): 859–877. MR3671776. doi: https://doi.org/10.1080/01621459.2017.1285773. 1509

Blei, D. M. and Lafferty, J. D. (2007). "A correlated topic model of Science." *The Annals of Applied Statistics*, 1(1): 17–35. MR2393839. doi: https://doi.org/10.1214/07-AOAS114. 1494, 1502, 1504, 1509

Boogaart, K. and Tolosana-Delgado, R. (2013). *Analyzing Compositional Data with R*, 209–253. MR3099409. doi: https://doi.org/10.1007/978-3-642-36809-7. 1502, 1504, 1505

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). "Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors." *Bayesian Analysis*, 1–30. MR4770323. doi: https://doi.org/10.1214/23-BA1371. 1518

Boulesteix, N. K. A.-L. (2014). *ppls: Penalized Partial Least Squares*. R package version 1.6-1. URL http://cran.nexr.com/web/packages/ppls/index.html 1516

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition. 1510, 1512

Bürkner, P.-C. (2017). "brms: An R Package for Bayesian Multilevel Models Using Stan." *Journal of Statistical Software*, 80(1): 1–28. URL https://www.jstatsoft.org/index.php/jss/article/view/v080i01 1501

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1): 1–32. URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01 1495, 1510

Carvalho, Polson, and Scott (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480. URL http://www.jstor.org/stable/25734098 MR2650751. doi: https://doi.org/10.1093/biomet/asq017. 1493, 1494, 1496, 1497, 1498, 1510

Castillo, I. and van der Vaart, A. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069–2101. MR3059077. doi: https://doi.org/10.1214/12-AOS1029. 1497

Catalina, A., Bürkner, P.-C., and Vehtari, A. (2020). "Projection Predictive Inference for Generalized Linear and Additive Multilevel Models." 1517

Chen, J., Zhu, J., Wang, Z., Zheng, X., and Zhang, B. (2013). "Scalable Inference for Logistic-Normal Topic Models." In Burges, C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. (eds.), *Advances in Neural Information Processing Systems*, volume 26. Curran Associates, Inc. URL https://proceedings.neurips.cc/paper_files/paper/2013/file/285f89b802bcb2651801455c86d78f2a-Paper.pdf 1494, 1502, 1509

Chow, D. D. K. (2022). "Schlömilch integrals and probability distributions on the simplex." URL https://arxiv.org/abs/2201.11013 1502

Connor, R. J. and Mosimann, J. E. (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution." *Journal of the American Statistical Association*, 64(325): 194–206. URL http://www.jstor.org/stable/2283728 MR0240895. 1502, 1503

Creus-Martí, I., Moya, A., and Santonja, F.-J. (2021). "A Dirichlet Autoregressive Model for the Analysis of Microbiota Time-Series Data." *Complex.*, 2021: 9951817:1–9951817:16. URL https://api.semanticscholar.org/CorpusID:237715479 1505

Efron, B. (2011). "Tweedie's Formula and Selection Bias." *Journal of the American Statistical Association*, 106(496): 1602–1614. PMID: 22505788. MR2896860. doi: https://doi.org/10.1198/jasa.2011.tm11181. 1497

Filzmoser, P. (2023). *chemometrics: Multivariate Statistical Analysis in Chemometrics*. R package version 1.4.4. URL https://CRAN.R-project.org/package=chemometrics 1515

Follett, L. and Yu, C. (2019). "Achieving parsimony in Bayesian vector autoregressions with the horseshoe prior." *Econometrics and Statistics*, 11: 130–144. URL https://www.sciencedirect.com/science/article/pii/S2452306219300036 MR3980254. doi: https://doi.org/10.1016/j.ecosta.2018.12.004. 1496

Frederic, P. and Lad, F. (2008). "Two Moments of the Logitnormal Distribution." *Communications in Statistics – Simulation and Computation*, 37(7): 1263–1269. MR2528273. doi: https://doi.org/10.1080/03610910801983178. 1505

Gelman, A. (2006a). "Prior Distributions for Variance Parameters in Hierarchical Models." *Bayesian Analysis*, 1. 1498

Gelman, A. (2006b). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515–534. MR2221284. doi: https://doi.org/10.1214/06-BA117A. 1501

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. URL https://books.google.de/books?id=ZXL6AQAAQBAJ MR3235677. 1493

Gelman, A. and Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press. MR2262986. doi: https://doi.org/10.1137/040607964. 1493

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press. 1497

Geyer, C. J. (1992). "Practical Markov Chain Monte Carlo." *Statistical Science*, 7(4): 473–483. doi: https://doi.org/10.1214/ss/1177011137 1512

Ghosal, S., Ghosh, J. K., and van der Vaart, A. W. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500–531. MR1790007. doi: https://doi.org/10.1214/aos/1016218228. 1494

Ghosh, J. and Ghattas, A. (2015). "Bayesian Variable Selection Under Collinearity." *The American Statistician*, 69. MR3391636. doi: https://doi.org/10.1080/00031305.2015.1031827. 1516

Ghosh, S., Yao, J., and Doshi-Velez, F. (2019). "Model Selection in Bayesian Neural Networks via Horseshoe Priors." *Journal of Machine Learning Research*, 20(182): 1–46. URL http://jmlr.org/papers/v20/19-236.html MR4048993. 1496

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. URL https://books.google.de/books?id=qRuVoAEACAAJ MR3307991. 1493, 1494, 1506

Good, I. J. (1962). "Theory of Probability Harold Jeffreys (Third edition, 447 + ix pp., Oxford Univ. Press, 84s.)." *Geophysical Journal International*, 6: 555–558. MR0187257. 1501

Goodfellow, I. J., Bengio, Y., and Courville, A. (2016). *Deep Learning*. Cambridge, MA, USA: MIT Press. http://www.deeplearningbook.org. MR3617773. 1504

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). "rstanarm: Bayesian applied regression modeling via Stan." R package version 2.21.1. URL https://mc-stan.org/rstanarm 1501

Greenacre, M., Grunsky, E., Bacon-Shone, J., Erb, I., and Quinn, T. (2023). "Aitchison's Compositional Data Analysis 40 Years on: A Reappraisal." *Statistical Science*, 38(3): 386–410. MR4630375. doi: https://doi.org/10.1214/22-STS880. 1494, 1502, 1504, 1505

Griffin, J. E. and Brown, P. J. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171–188. MR2596440. doi: https://doi.org/10.1214/10-BA507. 1494, 1496, 1510

Griffin, J. E. and Brown, P. J. (2013). "Some Priors for Sparse Regression Modelling." *Bayesian Analysis*, 8(3): 691–702. MR3102230. doi: https://doi.org/10.1214/13-BA827. 1510, 1516

Gupta, R. D. and Richards, D. S. P. (2001). "The History of the Dirichlet and Liouville Distributions." *International Statistical Review / Revue Internationale de Statistique*, 69(3): 433–446. URL http://www.jstor.org/stable/1403455 1502

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. ISSN. CRC Press. URL https://books.google.de/books?id=f-A_CQAAQBAJ MR3616141. 1494

Hintze, J. L. and Nelson, R. D. (1998). "Violin Plots: A Box Plot-Density Trace Synergism." *The American Statistician*, 52(2): 181–184. URL https://www.tandfonline.com/doi/abs/10.1080/00031305.1998.10480559 1513

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67. URL http://www.jstor.org/stable/1267351 MR0611894. doi: https://doi.org/10.1080/01966324.1981.10737061. 1493

Hoffman, M. D. and Gelman, A. (2011). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." 1510

Holmes, J. B. and Schofield, M. R. (2022). "Moments of the logit-normal distribution." *Communications in Statistics – Theory and Methods*, 51(3): 610–623. MR4368427. doi: https://doi.org/10.1080/03610926.2020.1752723. 1505

Jeffrey, A., Zwillinger, D., Gradshteyn, I., and Ryzhik, I. (2007). "8–9 – Special Functions." In *Table of Integrals, Series, and Products (Seventh Edition)*, 859–1048. Boston: Academic Press, seventh edition. URL https://www.sciencedirect.com/science/article/pii/B9780080471112500169 MR2360010. 1508

Johndrow, J., Orenstein, P., and Bhattacharya, A. (2020). "Scalable Approximate MCMC Algorithms for the Horseshoe Prior." *Journal of Machine Learning Research*, 21(73): 1–61. URL http://jmlr.org/papers/v21/19-536.html MR4095352. 1494, 1496

Kohns, D. and Szendrei, T. (2024). "Horseshoe prior Bayesian quantile regression." *Journal of the Royal Statistical Society Series C: Applied Statistics*, 73(1): 193–220. MR4719324. doi: https://doi.org/10.1093/jrsssc/qlad091. 1496

Kruijer, W., Rousseau, J., and Vaart, A. (2010). "Adaptive Bayesian Density Estima-

tion with Location-Scale Mixtures." *Electronic Journal of Statistics*, 4. MR2735885. doi: https://doi.org/10.1214/10-EJS584. 1500

Kruschke, J. K. (2015). "Chapter 6 – Inferring a Binomial Probability via Exact Mathematical Analysis." In Kruschke, J. K. (ed.), *Doing Bayesian Data Analysis (Second Edition)*, 123–141. Boston: Academic Press, second edition. URL https://www.sciencedirect.com/science/article/pii/B9780124058880000064 1500

Kullback, S. and Leibler, R. A. (1951). "On Information and Sufficiency." *The Annals of Mathematical Statistics*, 22(1): 79–86. MR0039968. doi: https://doi.org/10.1214/aoms/1177729694. 1508

Lin, J. (2016). "On The Dirichlet Distribution by Jiayu Lin." 1499, 1500, 1502

McLatchie, Y., Rögnvaldsson, S., Weber, F., and Vehtari, A. (2023). "Robust and efficient projection predictive inference." MR4859120. doi: https://doi.org/10.1214/24-sts949. 1517

Mikkola, P., Martin, O. A., Chandramouli, S., Hartmann, M., Pla, O. A., Thomas, O., Pesonen, H., Corander, J., Vehtari, A., Kaski, S., Bürkner, P.-C., and Klami, A. (2021). "Prior knowledge elicitation: The past, present, and future." 1495

Mimno, D., Wallach, H., and McCallum, A. (2008). "Gibbs sampling for logistic normal topic models with graph-based priors." 1509

Ongaro, A. and Migliorati, S. (2013). "A generalization of the Dirichlet distribution." *Journal of Multivariate Analysis*, 114: 412–426. URL https://www.sciencedirect.com/science/article/pii/S0047259X12001753 MR2993896. doi: https://doi.org/10.1016/j.jmva.2012.07.007. 1502, 1503

Osborne, B., Fearn, T., Miller, A. R., and Douglas, S. (1984). "Application of near infrared reflectance spectroscopy to the compositional analysis of biscuits and biscuit doughs." *Journal of the Science of Food and Agriculture*, 35: 99–105. URL https://api.semanticscholar.org/CorpusID:94286061 1515, 1516

Pavone, F., Piironen, J., Bürkner, P.-C., and Vehtari, A. (2020). "Using reference models in variable selection." 1497

Pawlowsky-Glahn, V. and Buccianti, A. (2011). "Compositional data analysis: theory and applications." URL https://api.semanticscholar.org/CorpusID:117956959 MR2920574. doi: https://doi.org/10.1002/9781119976462. 1503, 1504, 1505

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155–2197. MR4097052. doi: https://doi.org/10.1214/20-EJS1711. 1497, 1517

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018–5051. MR3738204. doi: https://doi.org/10.1214/17-EJS1337SI. 1494, 1496, 1497, 1517

Polson, N. G. and Scott, J. G. (2011). "Local shrinkage rules, Levy processes, and regularized regression." MR2899864. doi: https://doi.org/10.1111/j.1467-9868.2011.01015.x. 1516

Polson, N. G. and Scott, J. G. (2012). "On the Half-Cauchy Prior for a Global Scale Parameter." *Bayesian Analysis*, 7(4): 887–902. MR3000018. doi: https://doi.org/10.1214/12-BA730. 1498

Polson, N. G., Scott, J. G., and Windle, J. (2013). "Bayesian inference for logistic models using Pólya–Gamma latent variables." *Journal of the American statistical Association*, 108(504): 1339–1349. MR3174712. doi: https://doi.org/10.1080/01621459.2013.829001. 1497

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York. URL https://books.google.ch/books?id=6oQ4s8Pq9pYC MR2723361. 1512

Ročková, V. (2018). "Bayesian estimation of sparse signals with a continuous spike-and-slab prior." *Annals of Statistics*, 46: 401–437. URL https://api.semanticscholar.org/CorpusID:85503428 MR3766957. doi: https://doi.org/10.1214/17-AOS1554. 1494

Song, Q. and Liang, F. (2017). "Nearly optimal Bayesian shrinkage for high-dimensional regression." *Science China Mathematics*, 66: 409–442. URL https://api.semanticscholar.org/CorpusID:88516067 MR4535982. doi: https://doi.org/10.1007/s11425-020-1912-6. 1497

Stan Development Team (2024). "Stan Modeling Language Users Guide and Reference Manual, Version 2.36." URL http://mc-stan.org/ 1495, 1510

Stein, C. M. (1981). "Estimation of the Mean of a Multivariate Normal Distribution." *The Annals of Statistics*, 9(6): 1135–1151. Publisher: Institute of Mathematical Statistics. URL https://projecteuclid.org/journals/annals-of-statistics/volume-9/issue-6/Estimation-of-the-Mean-of-a-Multivariate-Normal-Distribution/10.1214/aos/1176345632.full MR0630098. 1496

Stryjewski, L. (2010). "40 years of boxplots." URL https://api.semanticscholar.org/CorpusID:36975036 1513

Stuart, A. and Ord, K. (2009). *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory*. Number vol. 1 ;vol. 1994 in Kendall's Advanced Theory of Statistics. Wiley. URL https://books.google.ch/books?id=tW18thQWJQIC MR1285356. 1506

Szakács, G., Annereau, J.-P., Lababidi, S., Shankavaram, U., Arciello, A., Bussey, K. J., Reinhold, W., Guo, Y., Kruh, G. D., Reimers, M., Weinstein, J. N., and Gottesman, M. M. (2004). "Predicting drug sensitivity and resistance: Profiling ABC transporter genes in cancer cells." *Cancer Cell*, 6(2): 129–137. URL https://www.sciencedirect.com/science/article/pii/S1535610804002065 1516

Tadesse, M. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. URL https://books.google.de/books?id=Cn1TEAAAQBAJ 1494, 1496, 1517

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. MR1379242. 1493, 1494

Tosh, C., Greengard, P., Goodrich, B., Gelman, A., Vehtari, A., and Hsu, D. (2022). "The piranha problem: Large effects swimming in a small pond." 1509

Van der Pas, S. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC. 1493, 1496, 1497

Van der Pas, S., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976–1000. MR3486423. doi: https://doi.org/10.1214/16-EJS1130. 1493, 1494, 1496, 1498

Van der Pas, S., Szabó, B., and van der Vaart, A. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221–1274. MR3724985. doi: https://doi.org/10.1214/17-BA1065. 1494, 1497

Van der Pas, S. L., Kleijn, B. J. K., and van der Vaart, A. W. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585–2618. MR3285877. doi: https://doi.org/10.1214/14-EJS962. 1494

Van Erp, S., Oberski, D., and Mulder, J. (2019). "Shrinkage priors for Bayesian penalized regression." *Journal of Mathematical Psychology*, 89: 31–50. MR3903921. doi: https://doi.org/10.1016/j.jmp.2018.12.004. 1496, 1497

Vehtari, A., Gabry, J., Magnusson, M., Yao, Y., Bürkner, P.-C., Paananen, T., and Gelman, A. (2023). "loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models." R package version 2.6.0. URL https://mc-stan.org/loo/ 1516

Vehtari, A., Gelman, A., and Gabry, J. (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432. MR3647105. doi: https://doi.org/10.1007/s11222-016-9696-4. 1512, 1516

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667–718. MR4298989. doi: https://doi.org/10.1214/20-BA1221. 1512

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142–228. MR3011074. doi: https://doi.org/10.1214/12-SS102. 1512

Vehtari, A., Simpson, D., Gelman, A., Yao, Y., and Gabry, J. (2022). "Pareto Smoothed Importance Sampling." MR4749108. 1512

Vershynin, R. (2018). *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge Series in Statistical and Probabilistic Mathematics. Cam-

bridge University Press. URL https://books.google.de/books?id=J-VjswEACAAJ MR3837109. doi: https://doi.org/10.1017/9781108231596. 1506

Wang, Y. and Polson, N. G. (2024). "Pochhammer Priors for Sparse Count Models." URL https://arxiv.org/abs/2402.09583 1502

West, M. (1987). "On scale mixtures of normal distributions." *Biometrika*, 74(3): 646–648. MR0909372. doi: https://doi.org/10.1093/biomet/74.3.646. 1496, 1501

Wong, T.-T. (1998). "Generalized Dirichlet distribution in Bayesian analysis." *Applied Mathematics and Computation*, 97(2): 165–181. URL https://www.sciencedirect.com/science/article/pii/S0096300397101400 MR1643091. doi: https://doi.org/10.1016/S0096-3003(97)10140-0. 1503

Xun, G., Li, Y., Zhao, W. X., Gao, J., and Zhang, A. (2017). "A Correlated Topic Model Using Word Embeddings." In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 4207–4213. doi: https://doi.org/10.24963/ijcai.2017/588 1509

Yang, Y., Wainwright, M. J., and Jordan, M. I. (2016). "On the Computational Complexity of High-Dimensional Bayesian Variable Selection." *The Annals of Statistics*, 44(6): 2497–2532. URL http://www.jstor.org/stable/44245760 MR3576552. doi: https://doi.org/10.1214/15-AOS1417. 1494

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823–844. doi: https://doi.org/10.1214/17-BA1076 1494, 1497, 1499, 1517

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13. doi: https://doi.org/10.1080/01621459.2020.1825449 1494, 1495, 1496, 1499, 1501, 1505, 1510, 1516, 1517

Generalized Decomposition Priors on R2

Javier Enrique Aguilar [*,†,‡] and Paul-Christian Bürkner[*,§]

S1 Supplementary Results

S1.1 Shifting between Logistic Normal representations

We have explored different transformations from the real space into the simplex. Namely the additive log-ratio (alr) and the symmetric logistic transformation. The alr transformation relative to the Kth component is defined as: (Aitchison and Shen, 1980).

$$\eta = \text{alr}(\phi) := \left(\ln\left(\frac{\phi_1}{\phi_K}\right), ..., \ln\left(\frac{\phi_{K-1}}{\phi_K}\right) \right) \in \mathbb{R}^{K-1}, \tag{1}$$

where $\phi \in \mathcal{S}^K$. The inverse alr transformation is defined by the *standard logistic* (*softmax*) transformation with a fill-up term for the reference component: $\phi_k = \frac{e^{\eta_k}}{1+\sum_{j=1}^{K-1} e^{\eta_j}}, k = 1, ..., K-1, , \phi_K = 1 - \sum_{k=1}^{K-1} \phi_k$.

If $\eta = \text{alr}(\phi)$ follows a multivariate normal distribution in $\mathbb{R}^{K-1}$, then ϕ is said to follow an additive logistic normal (ALN) distribution, denoted as $\phi \sim \text{ALR}(\mu, \Sigma)$. The parameters μ and Σ depend on the chosen reference component.

The symmetric logistic transformation does not rely on a reference component. Here, we start with an unconstrained vector $\theta \in \mathbb{R}^K$, assumed to follow a multivariate normal distribution $\theta \sim \mathcal{N}_K(\mu, \Sigma)$, and map it onto the simplex via

$$\phi_k = \frac{e^{\theta_k}}{\sum_{j=1}^{K} e^{\theta_j}}, \ k = 1, \ldots, K. \tag{2}$$

In this case, ϕ follows a logistic normal (LN) distribution, written as $\phi \sim \text{LogisticNormal}(\mu, \Sigma)$. To ensure identifiability, it is common to impose either the constraint $\sum_{k=1}^{K} \theta_k = 0$ or fix one component, e.g., $\theta_K = 0$. Other options are available as well. The latter case corresponds to the ALR transformation, while under the sum-to-zero constraint, we can express θ_k in terms of ϕ as $\theta_k = \ln\left(\frac{\phi_k}{g(\phi)}\right)$, where $g(\phi)$ is the geometric mean of ϕ.

To translate between the two representations, we consider their associated mean and covariance structures. Let $\Sigma = [\sigma_{ij}]$ denote the covariance matrix of the alr-transformed variables,

$$\sigma_{ij} = \text{cov}\left(\log\left(\frac{\phi_i}{\phi_K}\right), \log\left(\frac{\phi_j}{\phi_K}\right) \right), \quad i, j = 1, \ldots, K-1.$$

[*]Department of Statistics, TU Dortmund University, Germany
[†]Cluster of Excellence SimTech, University of Stuttgart, Germany
[‡]Corresponding author: javier.aguilarr@icloud.com url: https://jear2412.github.io
[§]paul.buerkner@gmail.com url: https://paul-buerkner.github.io

and let $\Gamma = [\gamma_{ij}]$ be the covariance matrix of the centered log-transformed variables,

$$\gamma_{ij} = \text{cov}\left(\log\left(\frac{\phi_i}{g(\phi)}\right), \log\left(\frac{\phi_j}{g(\phi)}\right)\right), \quad i,j = 1,\ldots,K.$$

The expectation and covariance relationships between η and θ are then given by

$$\begin{aligned}
\mathbb{E}(\eta_i) &= \mathbb{E}(\theta_i) - \mathbb{E}(\theta_K), \quad i = 1,\ldots,K-1,\\
\mathbb{E}(\theta_i) &= \frac{K-1}{K}\mathbb{E}(\eta_i) - \sum_{j\neq i}\mathbb{E}(\eta_j), \quad i = 1,\ldots,K-1,\\
\sigma_{ij} &= \gamma_{ij} - \gamma_{iK} - \gamma_{jK} + \gamma_{KK}, \quad i,j = 1,...,K-1\\
\gamma_{ij} &= \sigma_{ij} - \sigma_{i\cdot} - \sigma_{j\cdot} + \sigma_{\cdot\cdot} \quad i,j = 1,...,K,
\end{aligned}$$

where the subscript $\cdot$ denotes summation over all possible values of that index, followed by division by K. For instance $\sigma_{i\cdot} = K^{-1}\sum_j \sigma_{ij}$ and $\sigma_{\cdot\cdot} = K^{-2}\sum_i\sum_j \sigma_{ij}$.

The change of representation is particularly useful when employing prior matching, where the user first determines μ^*, Σ^* for η and then derives the corresponding mean and covariance Γ for θ.

S1.2 Taylor approximation for the mean of a ratio

The following result appears in Stuart and Ord (2009). We show it, given it is central to the discussion of hyperparameter specification.

Let X and Y have support in $(0,\infty)$. For $f(x,y)$, the bivariate first order Taylor expansion around (a,b) is given by

$$f(x,y) = f(a,b) + f'_x(a,b)(x-a) + f'_y(a,b)(y-a) + Q$$

where f_x, f_y denote the partial derivatives and Q is a remainder of order smaller than the other terms. Assume that $\mu_x = \mathbb{E}(X), \mu_y = \mathbb{E}(Y)$ exist and select the point of expansion as (μ_x,μ_y). An approximation to $\mathbb{E}(f(X,Y))$ is given by

$$\begin{aligned}
\mathbb{E}\left(f(X,Y)\right) &\approx \mathbb{E}\left[f\left(\mu_x,\mu_y\right) + f'_x\left(\mu_x,\mu_y\right)\left(X-\mu_x\right) + f'_y\left(\mu_x,\mu_y\right)\left(Y-\mu_y\right)\right]\\
&= \mathbb{E}\left[f\left(\mu_x,\mu_y\right)\right] + f'_x\left(\mu_x,\mu_y\right)\mathbb{E}\left[\left(X-\mu_x\right)\right] + f'_y\left(\mu_x,\mu_y\right)\mathbb{E}\left[\left(Y-\mu_y\right)\right]\\
&= f(\mu_x,\mu_y).
\end{aligned}$$

If $f(x,y)) = \frac{x}{y}$, then using (μ_x,μ_y) as expansion point, results in $\mathbb{E}\left(\frac{X}{Y}\right) = \frac{\mu_x}{\mu_y}$.

We can improve the first order approximation. The second order Taylor approximation around (a, b) is given by

$$f(x,y) = f(a,b) + f'_x(a,b)(x-a) + f'_y(a,b)(y-a) + \\ 1/2\left[f''_{xx}(x,y)(x-a)^2 + 2f''_{xy}(x,y)(x-a)(y-b) + f''_{yy}(x,y)(y-b)^2\right] + Q,$$

where f_{xx}, f_{yy} are the second order partial derivatives and Q is of small order. Choosing $(a, b) = (\mu_x, \mu_y)$ yields

$$\mathbb{E}\left(f(X,Y)\right) \approx f(\mu_x,\mu_y) + 1/2\left[f''_{xx}(\mu_x,\mu_y)\text{var}(X) + 2f''_{xy}(\mu_x,\mu_y)\text{Cov}(X,Y) + f''_{yy}(\mu_x,\mu_y)\text{var}(Y)\right].$$

If $f(x,y)) = \frac{x}{y}$, then using (μ_x, μ_y) as expansion point, provides us with $\mathbb{E}\left(\frac{X}{Y}\right) \approx \frac{\mathbb{E}[X]}{\mathbb{E}[Y]} - \frac{\text{Cov}(X,Y)}{(\mathbb{E}[Y])^2} + \frac{\text{Var}(Y)\cdot\mathbb{E}[X]}{(\mathbb{E}[Y])^3}$. The last expression is not always available in close form and we must resort to Monte Carlo methods.

S1.3 Gibbs with slice sampling

A Gibbs sampler for the case in which a Dirichlet decomposition is involved has been proposed by Aguilar and Bürkner (2023). We proceed to show a Gibbs sampler with an imbedded slice sampler for the case in which a Logistic Normal decomposition is used (Neal, 2003; Robert and Casella, 2004). For ease of notation and to be able to develop a blocked Gibbs sampling we will write the GDR2 model in matrix form. This will drastically improve performance over moving one regression term at a time.

Denote by $b = (b_1, ..., b_K)'$ and let $y = (y_1, ..., y_N)'$ denote a vector containing the N observations $y_i, i = 1, ..., N$, X denote the design matrix of dimension $N \times p$. Let $\Sigma_b = \sigma^2\Gamma_b$ denote the covariance matrix of the vector of the coefficients b where $\Gamma_b = \text{diag}\left\{\phi_1\tau^2, ..., \phi_K\tau^2\right\}$. Let I_N represent the identity matrix of order N. With this notation we can write the GDR2 model in matrix form as:

$$\begin{aligned} y \mid \mu, \sigma^2 &\sim \mathcal{N}(\mu, \sigma^2 I_N), \\ b \mid \sigma, \tau^2, \phi &\sim \mathcal{N}\left(0, \Sigma_b\right), \\ \phi &\sim \text{LN}(\mu, \Sigma) \\ \tau^2 \mid \xi \sim \text{Gamma}(a_1, \xi), &\quad \xi \sim \text{Gamma}(a_2, 1), \quad \sigma \sim p(\sigma). \end{aligned} \tag{3}$$

We have used the fact that $\tau^2 \mid \xi \sim \text{Gamma}(a_1, \xi), \xi \sim \text{Gamma}(a_2, 1)$ is equivalent to $\tau^2 \sim \text{BetaPrime}(a_1, a_2)$ (Zhang et al., 2020). In the following we denote by $Z \sim \text{GenInvGaussian}\left(\chi, \rho, \nu\right)$, the Generalized Inverse Gaussian distribution (Robert, 1991) with parameters $\chi > 0, \rho > 0$ and $\nu \in \mathbb{R}$ if

$$p(z) \propto z^{\nu-1}\exp\left\{-\left(\rho z + \chi/z\right)/2\right\}.$$

We denote by InvGamma(c, d) the Inverse Gamma distribution with shape c and scale d and consider that a priori $\sigma^2 \sim \text{InvGamma}(c, d)$. Consider representation (3) of the GDR2 prior, then the Gibbs sampling procedure is the following:

1. Set initial values for $b, \sigma, \phi, \tau^2, \xi$.
2. Sample $b \mid \phi, \tau, \sigma, y \sim \mathcal{N}\left(\bar{b}, S_b\right)$, where

$$\bar{b} = \left(X'X + \Gamma_b^{-1}\right)^{-1} X'y, \quad S_b = \sigma^2 \left(X'X + \Gamma_b^{-1}\right)^{-1}.$$

3. Sample $\sigma^2 \mid b, \phi, \tau^2 \sim \text{InvGamma}\left(\dot{c}, \dot{d}\right)$, where

$$\dot{c} = c + \frac{1}{2}\left(N + K\right), \quad \dot{d} = d + \frac{1}{2}\left(||y - Xb||_2^2 + b'\Gamma_b^{-1}b\right).$$

4. Sample $\tau^2 \mid b, \sigma, \xi \sim \text{GenInvGaussian}\left(\chi, \rho, \nu\right)$, where

$$\chi = \frac{1}{\sigma^2} b' T_b^{-1} b, \quad \rho = 2\xi, \quad \nu = a_1 - K/2,$$

   where T_b is such that $\Gamma_b = T_b \tau^2$. Sampling from the Generalized Inverse Gaussian is not straightforward, but the reader can refer to Robert (1991); Leydold and Hörmann (2011) for efficient methods.

5. Sample $\xi \mid \tau^2 \sim \text{Gamma}\left(a_1 + a_2, 1 + \tau^2\right)$.

6. Sample $\phi \mid b, u, \sigma, \xi$.

   Instead of sampling from ϕ, we consider sampling from the vector of log ratios η. The conditional distribution of η is

$$p(\eta \mid b, \tau^2, \sigma) \propto \prod_{i=1}^{K} \exp\left\{-\frac{b_i^2}{2\tau^2\sigma^2}(1 + \sum_{j \neq i} \exp(\eta_j))\right\} \exp\left\{-1/2(\eta - \mu)'\Sigma^{-1}(\eta - \mu)\right\}. \tag{4}$$

   Follow the next steps to use slice sampling (Neal, 2003; Robert and Casella, 2004)

(a) The log-posterior is:

$$\log p(\eta|b,\tau^2,\sigma) = -\frac{1}{2}(\eta-\mu)'\Sigma^{-1}(\eta-\mu) - \sum_{i=1}^{K}\frac{b_i^2}{2\tau^2\sigma^2}\left(1+\sum_{j\neq i}\exp(\eta_j)\right).$$

(b) Use slice sampling for each dimension η_i. For each dimension of $\eta = (\eta_1, \eta_2, \ldots, \eta_K)$ do the following:

i. Introduce an auxiliary variable u. Sample u uniformly from $[0, p(\eta \mid b, \tau^2, \sigma)$. This defines the slice:

$$u \leq \log p(\eta|b,\tau^2,\sigma).$$

ii. Create an initial interval $[L, R]$ around the current value of η_i:

$$L = \eta_i - w, \quad R = \eta_i + w,$$

where w is a tuning parameter.

iii. Propose a new value for η_i uniformly from $[L, R]$ and accept it only if $\log p(\eta) \geq u$.

iv. If the proposed value is rejected, shrink the interval $[L, R]$ and repeat until a valid sample is found.

v. Replace the current η_i with the accepted sample and proceed to the next dimension.

vi. Repeat the process for all dimensions to complete the slice sampling step.

7. Repeat until convergence

It is also possible to create a multivariate slice sampler. See Neal (2003) for details.

S1.4 Stan code

The Stan codes for the GDR2 prior, with its several decompositions, and the other priors used in the experiments are available at https://osf.io/ns2cv/, where it will be maintained and updated as needed.

S2 Further Simulation results

S2.1 GDR2 models

In the following we show additional simulation results for the simulated and fixed coefficient setup. We show Case (A) of the Simulated coefficient setup, in which each coefficient b_k is generated independently from a Normal distribution with mean 0 and variance $\sigma_b^2 = 9$. We also show the case in which the Fixed coefficients are set to a value of 7. For brevity, we show the cases in which $a_\pi = 0.5$. and $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$. The data for the other results can be found in https://osf.io/ns2cv/.

Results: Simulated coefficients setup

Figures 1 and 2 clearly demonstrate that, in the simpler scenario where no correlation is assumed among the components of b, LN decompositions outperform the Dirichlet decomposition. These results are consistent with, and even more promising than, those presented in the main text. However, the correlated case discussed in the main text is more representative of real-world applications and provides a more rigorous test of the priors' performance.

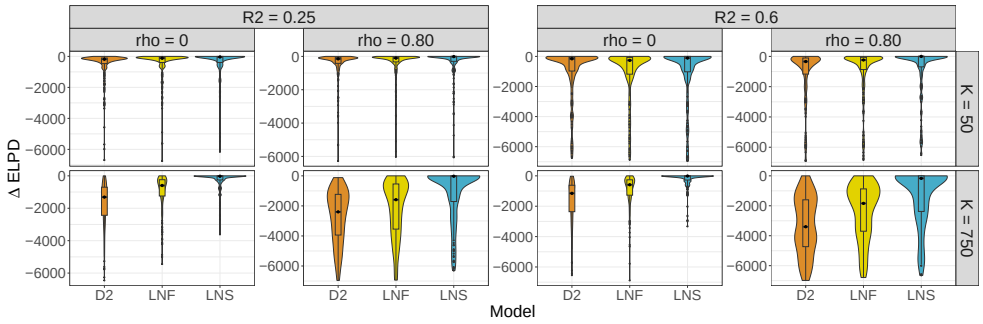


Figure 1: **Simulated coefficients setup.** ΔELPD evaluated on test datasets of size $N = 500$ for the simulated coefficients setup.

Results: Fixed coefficients setup

The results presented in Figures 3 and 4 align with the findings in the main text. In terms of predictive performance, the LNS consistently outperforms all other models. Regarding RMSE, the LN decompositions generally surpass the Dirichlet decompositions, except in the low-signal, high-correlation scenario for the nonzero coefficients. Notably, we observe improved performance when using the LNS instead of the LNF, supporting our earlier conjecture that an optimized choice of hyperparameters could further enhance performance. This also shows that a diagonal structure for Σ should serve as a default.

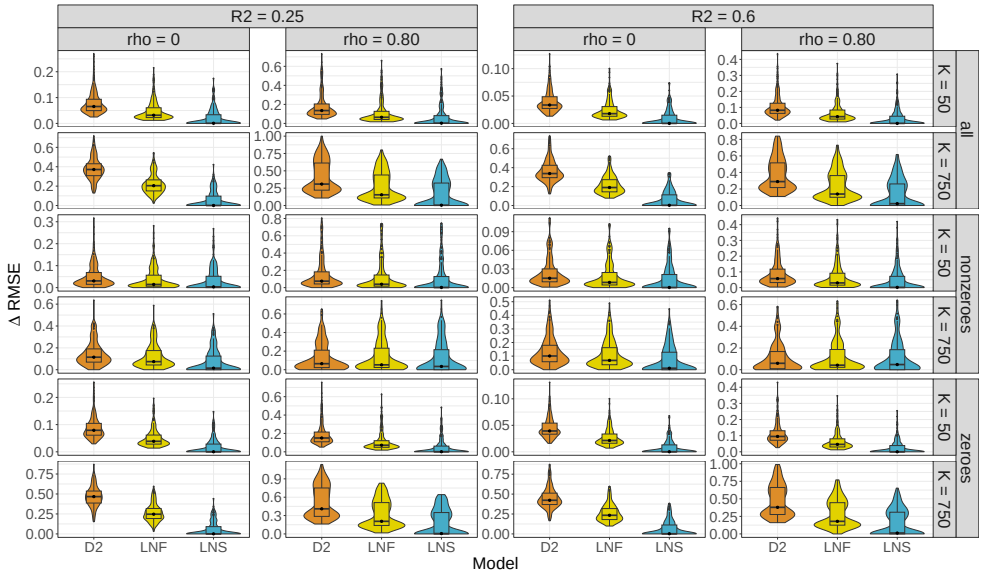


Figure 2: **Simulated coefficients setup.** Violin plots with embedded box plots for ΔRMSE under different simulation conditions. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

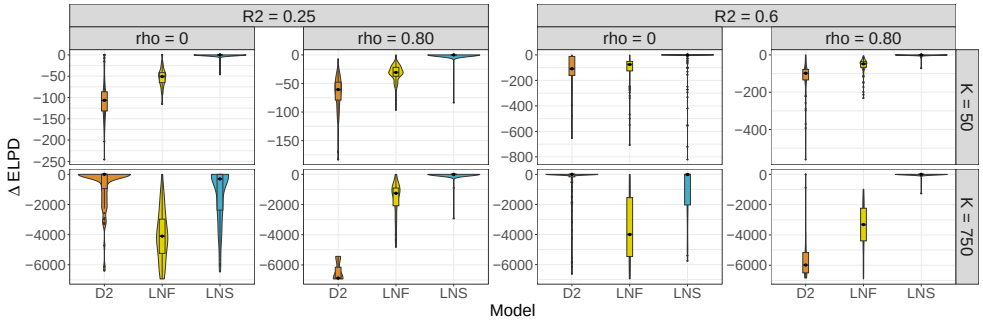


Figure 3: **Fixed coefficients setup.** ΔELPD evaluated on the test datasets for the fixed coefficients when the signals have a value of 7.

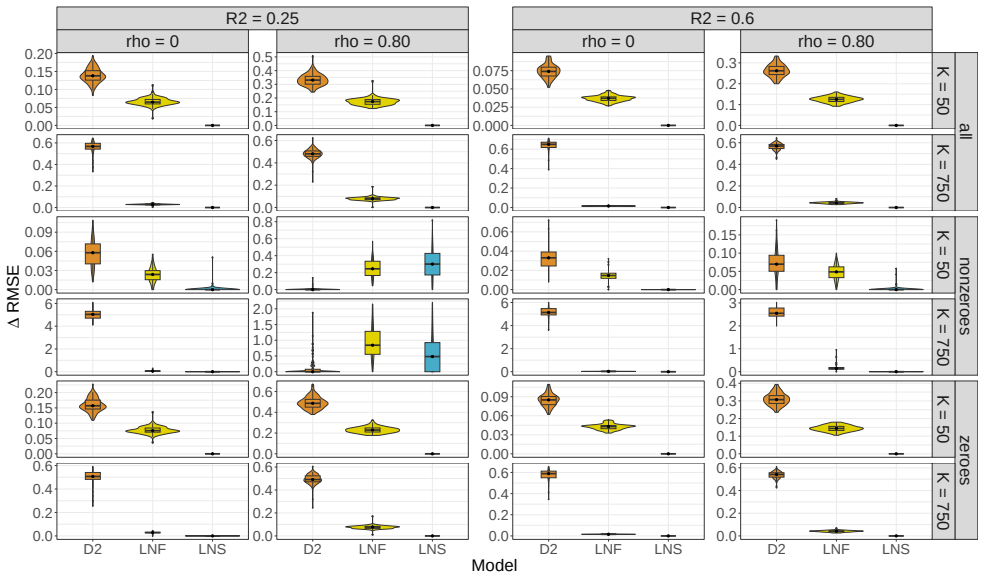


Figure 4: **Fixed coefficients setup.** Violin plots with embedded box plots for ΔRMSE under simulation conditions. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

Results: Fixed coefficients setup. Informed scenario.

In the fixed coefficient setup, we consider an extra condition, which we have added for the parameter α as follows: α_k takes the value of 10 when the true coefficient b_k is truly nonzero, and 0.5 when b_k is truly equal to zero. This reflects a scenario termed *informative* α, where strong user knowledge is incorporated to emphasize important signals. Our primary goals in examining this case are twofold: 1) Assessing the extent to which the Dirichlet and logistic normal decompositions benefit from user-provided information. 2) Investigating whether valuable information is effectively transferred from one prior to another through KL matching. While Dirichlet distributions have been employed in hierarchical priors previously (Aguilar and Bürkner, 2023), our search did not yield a study where information is directly injected into hyperparameters as demonstrated in this case. Other researchers have highlighted the significance of exploring hyperparameter values that imbue the prior with specific properties (Bhattacharya et al., 2015; Zhang et al., 2020; Aguilar and Bürkner, 2023). It is important to note that this informative scenario described here is uncommon in practice, particularly in high-dimensional settings where such detailed prior knowledge is usually unavailable.

Figure 5 illustrates how the different models improve their predictive performance when properly guided by the true signals. Similarly, Figure 6 displays the distribution of ΔRMSE across various simulation conditions for the informed scenario. A closer examination reveals that both LN decompositions consistently outperform the Dirichlet decompositions, with the LNS emerging as the top-performing model.

In general, the three considered priors are becoming more similar, with no clear advantage for either prior. Keep in mind that, in high-dimensional real-world scenarios, users will almost never have this amount of prior knowledge available, rendering the informed prior scenario unrealistic in practice.

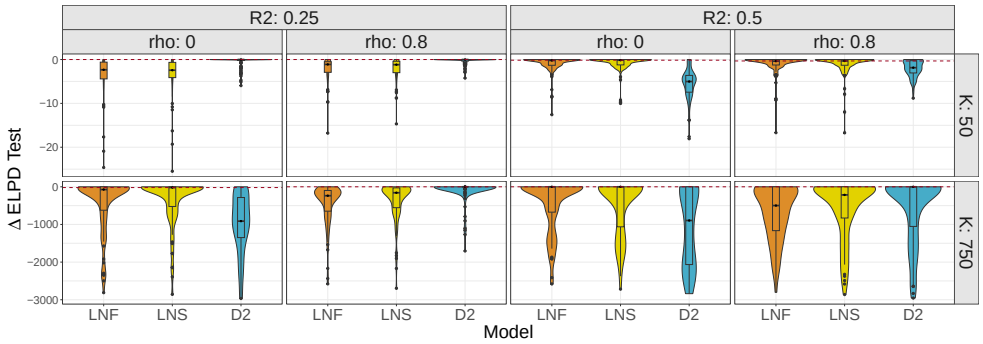


Figure 5: **Fixed coefficients setup. Informed scenario.** Violin plots with embedded box plots for ΔELPD for different simulation conditions.

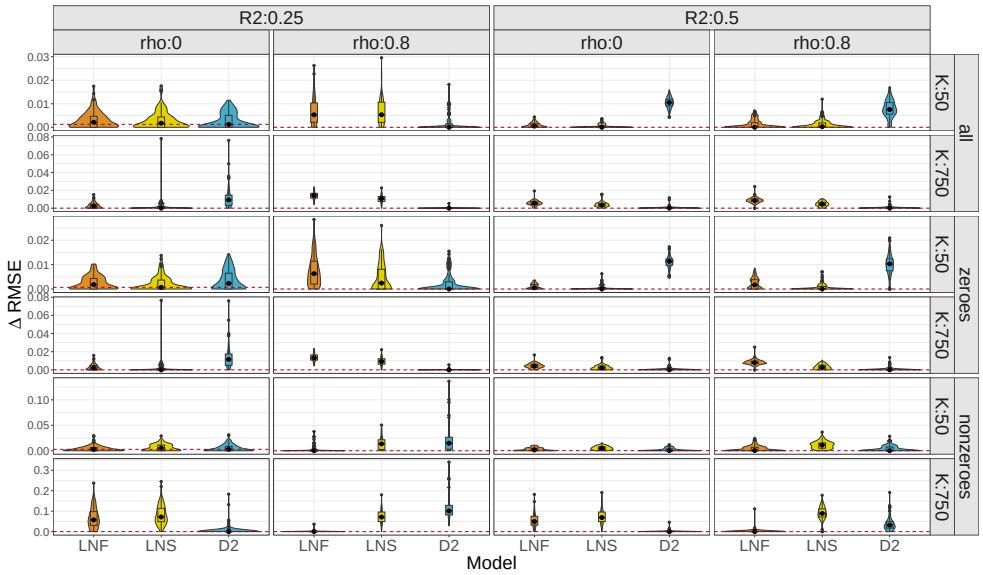


Figure 6: **Fixed coefficients setup. Informed scenario.** Violin plots with embedded box plots for ΔRMSE under different simulation conditions. The horizontal line shows the location of the minimum median per condition. We also provide a partition with respect to truly zero and nonzero coefficients.

S2.2 Comparison with other priors

The shrinkage priors analyzed in this study include the Beta Prime (BP), Dirichlet-Laplace (DL), Horseshoe (HS), and the proposed GDR2 model, which uses Dirichlet and Logistic Normal (LN) decompositions (Bai and Ghosh, 2019; Bhattacharya et al., 2015; Carvalho et al., 2010; Zhang et al., 2020). The Beta Prime, Dirichlet Laplace, and Horseshoe priors were implemented using default hyperparameters, which ensure concentrated mass near the origin and heavy tails. For the Simulated coefficients setup, we present results for a scenario where the coefficients follow an AR(1) correlation structure, defined by $\rho_b = 0.8$, marginal variances $\sigma_b^2 = 9$, and an average sparsity of $\nu = 0.75$. In the Fixed coefficients setup, the coefficients are assigned a constant value of 3. For brevity, we show the cases in which $a_\pi = 0.5$. and $(\mu_{R^2}, \varphi_{R^2}) = (0.5, 1)$. Additional results and analyses are available in `https://osf.io/ns2cv/`.

Results: Simulated coefficients setup

Figure 7 shows the distribution of ΔELPD across simulations under various conditions, visualized as violin plots with embedded boxplots. The results demonstrate that Logistic Normal (LN) decompositions consistently outperform their Dirichlet counterparts. Notably, the LN priors exhibit the best average performance across all considered scenarios. The GDR2 priors perform on par with other well-established shrinkage priors, outperforming them in several cases. In terms of predictive performance, the Beta Prime (BP) and Dirichlet Laplace (DL) priors perform poorly in high-dimensional settings, with values concentrating below $-1e5$ while the classical Horseshoe (HS) often achieves the best performance. Adjusting hyperparameters for R^2-based priors might bring them closer to the Horseshoe's performance; however, our results show that with sensible default settings, they perform competitively.

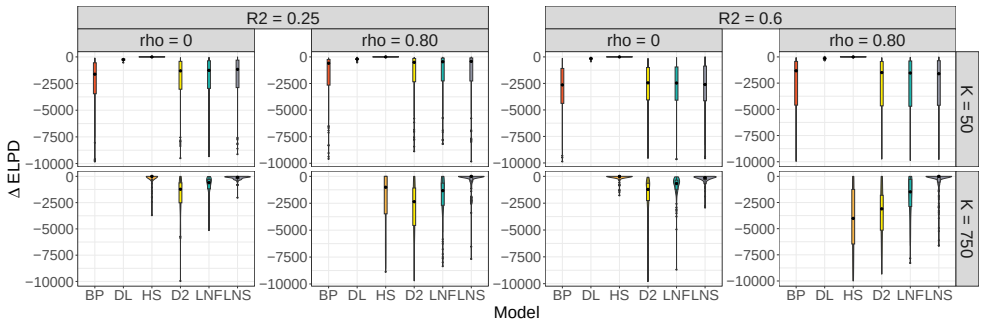


Figure 7: **Simulated coefficients setup:** ΔELPD evaluated on test datasets of size $N = 500$ for the simulated coefficients setup.

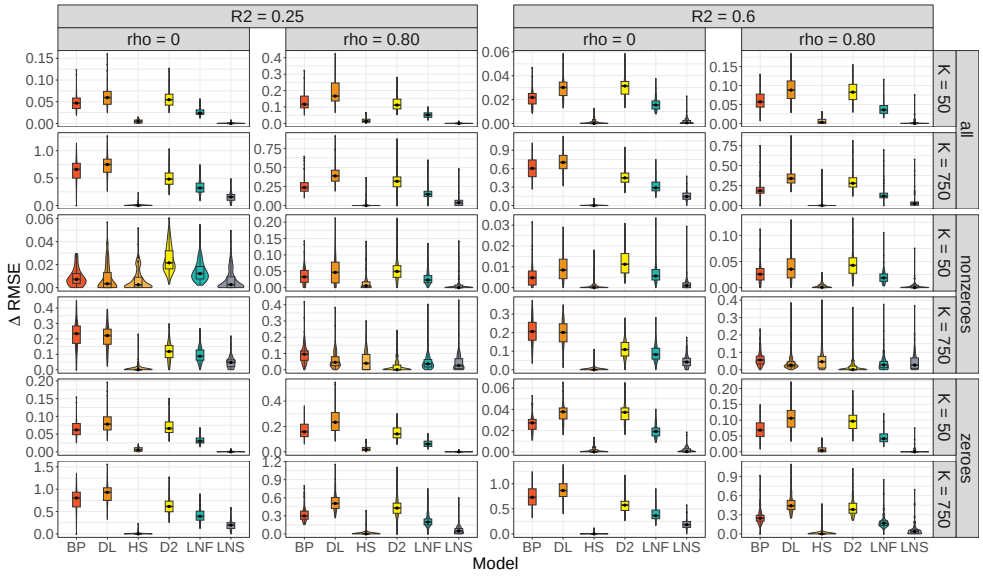


Figure 8: **Simulated coefficients setup:** Violin plots with embedded box plots for ΔRMSE under different simulation conditions. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

Figure 8 depicts the distribution of ΔRMSE across various simulation conditions. Similar to predictive performance, LN decompositions consistently outperform Dirichlet decompositions in terms of parameter recovery. The LN priors demonstrate the most robust parameter recovery among the competing R^2-based models, regardless of the values of ρ, R^2, or K. The overall RMSE results reflect a balance between the model's posterior error for truly zero and nonzero coefficients. To explore this further, Figure 8 separately illustrates ΔRMSE for truly zero and nonzero coefficients.

These results suggest that LN decompositions achieve more effective shrinkage and compared to Dirichlet counterparts, reducing false positives and improving the detection of truly zero coefficients. The trends for nonzero coefficients are similar. Of particular note is the case with $K = 750$, a challenging scenario where shrinkage priors are expected to overshrink. In this high-dimensional setting, R^2-based priors—especially LN decompositions—show improved behavior when moving from $\rho = 0$ to $\rho = 0.80$, performing on par even with the HS. This improvement is also present when transitioning from a low signal ($R^2 = 0.25$) to a moderate signal ($R^2 = 0.60$). The LN decomposition priors perform on par and improving over the performance of well established shrinkage priors.

We present coverage properties using 95% marginal credibility intervals in Table 1 for the case where $\rho = 0$ and $R^2 = 0.6$. Other results can be found in our repository https://osf.io/ns2cv/. The table reports the proportion of coverage, specificity, sensitivity, and the average width of the intervals. While frequentist properties are not

guaranteed in a Bayesian framework, all models achieve the nominal 95% coverage when $(K, N) = (50, 100)$ and $K = 150$, the latter case being impressive since we are in the high-dimensional setting. However, as K increases to 750, coverage declines due to the increased complexity of the problem.

Specificity and sensitivity remain high for $K = 50, 150$, but for $K = 750$, maintaining adequate sensitivity requires wider intervals, indicating greater uncertainty. Notably, the HS and LN models produce the shortest intervals in this setting. As expected, sensitivity declines sharply due to the increased interval width. Therefore, we recommend supplementing model fit with a second-stage procedure, such as Projection Prediction Piironen et al. (2020), for variable selection rather than relying solely on credibility intervals.

We present ROC curves in Figure 9, where the diagonal represents the identity line. The LN models perform comparably to the Horseshoe. In the complex setting with $K = 750$ and $\rho = 0$, the Horseshoe generally outperforms other models. However, when correlation is introduced in X, the LNS decomposition surpasses all alternatives.

Table 1: **Simulated coefficients setup:** Coverage properties of 95% marginal posterior credible intervals for the scenario $\rho = 0$, $R^2 = 0.6$

K	Model ID	Coverage	Specificity	Sensitivity (Power)	Avg. CI Width
50	BP	0.946	0.975	0.926	0.396
	DL	0.963	0.971	0.936	0.410
	HS	0.974	0.985	0.931	0.382
	D2	0.953	0.981	0.921	0.426
	LNF	0.958	0.988	0.923	0.406
	LNS	0.963	0.994	0.919	0.389
150	BP	0.978	0.997	0.707	1.500
	DL	0.993	1.000	0.592	2.212
	HS	0.971	0.987	0.758	1.189
	D2	0.985	1.000	0.584	2.192
	LNF	0.982	0.999	0.698	1.536
	LNS	0.978	0.998	0.748	1.205
750	BP	0.932	1.000	0.001	5.883
	DL	0.943	1.000	0.001	6.137
	HS	0.815	1.000	0.001	2.153
	D2	0.913	1.000	0.000	4.967
	LNF	0.890	1.000	0.000	4.311
	LNS	0.857	1.000	0.000	3.304

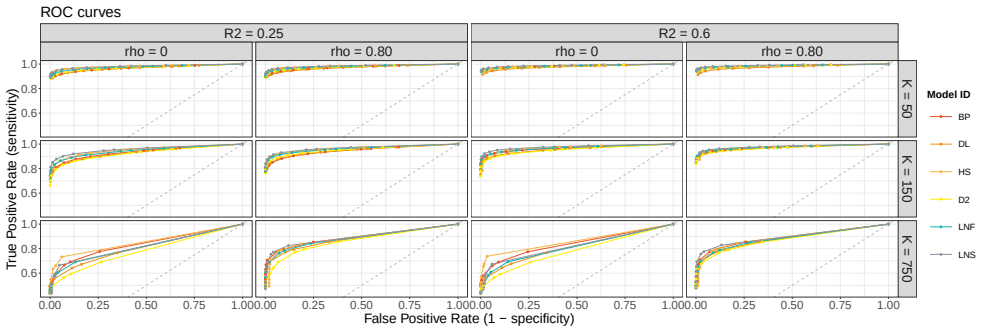


Figure 9: **Simulated coefficients setup:** ROC curves for different models under different simulation conditions. Points are obtained by varying the credibility level of marginal posterior credibility intervals. The grey dotted line depicts the identity line.

Figures 10, 11, and 12 present the Effective Sample Size (ESS), Rhat, and total runtime results across the simulations for the coefficient chains, respectively (Vehtari et al., 2021). The Rhat diagnostic indicates no significant convergence issues for most models, with the exception of the Horseshoe, where Rhat values deviate furthest from 1.0. ESS values are generally within acceptable ranges, suggesting efficient sampling across models, although the Horseshoe exhibits the lowest ESS, particularly as model complexity increases. Regarding runtime, Logistic-Normal (LN) models are the fastest for low-complexity cases; however, their computational time grows in higher-complexity scenarios. Finally, the effect of correlations is evident in the runtime distributions, with a noticeable increase in their medians as correlation strength increases.

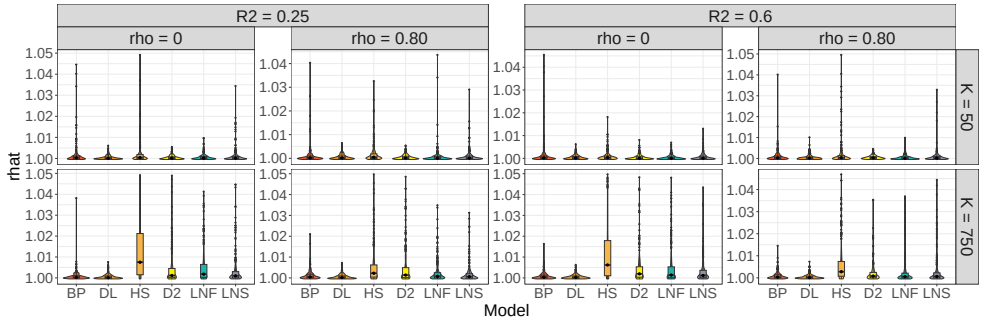


Figure 10: **Simulated coefficients setup:** Distribution of Rhat across different simulations.

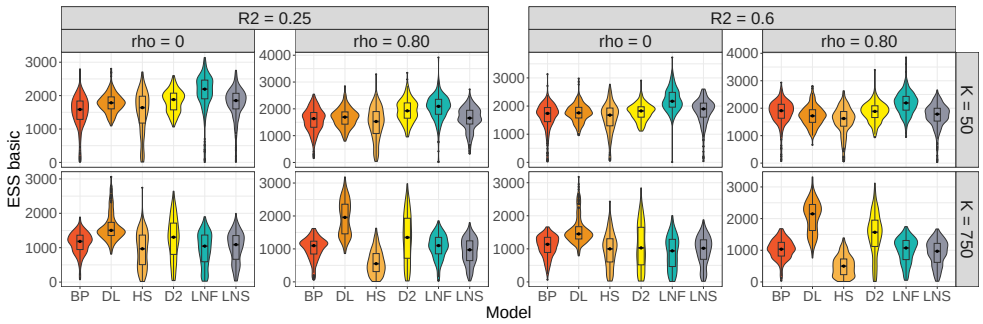


Figure 11: **Simulated coefficients setup:** Distribution of Effective Sample Size across different simulations.

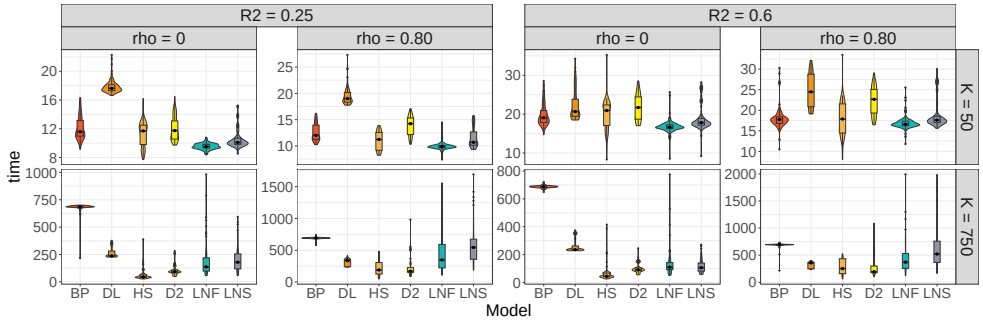


Figure 12: **Simulated coefficients setup:** Distribution of total sampling time for each model.

Results: Fixed coefficients setup

Figure 13 reveals that the LN decompositions are able to perform at par in out-of-sample predictive performance whenever X is uncorrelated and to outperform them whenever X is highly correlated. Among the R2-based priors, LNS stands out as the most effective predictive model across scenarios. The most pronounced disparities emerge in scenarios characterized by high correlation and dimensionality ($\rho = 0.8, K = 750$), where the performance of the Dirichlet prior is notably subpar. The values of BP are extremeley low, therefore not being considered in the plot.

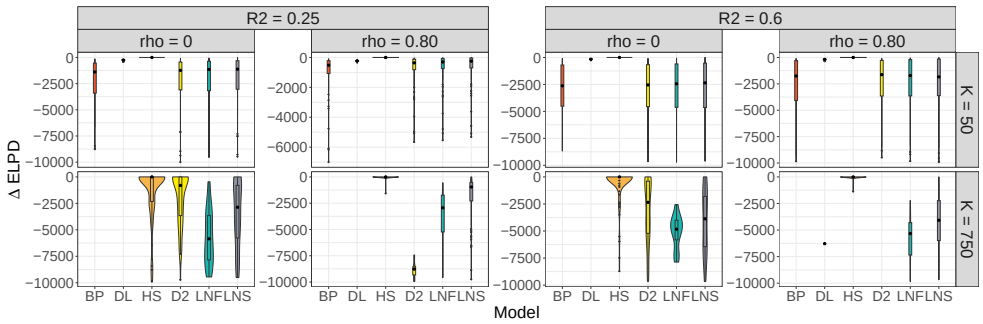


Figure 13: **Fixed coefficients setup:** ΔELPD evaluated on test datasets of size $N = 500$ for the simulated coefficients setup.

The results for ΔRMSE for all of the coefficients and the zero coefficients are very similar to the simulated coefficients setup. The LN decompositions perform clearly and uniformly better the Dirichlet for all the coefficients and for the zero coefficient only. This indicates again that LN decompositions are better at both overall RMSE and noise detection.

For ΔRMSE of nonzero coefficients, the Horseshoe tends to outperform other models across most scenarios, noenetheless LN decompositions perform on par. Among the R^2-based priors, no single model consistently dominates. The Dirichlet decomposition excels in finding nonzero coefficients in the in low-signal, high-correlation settings. LN decompositions achieves better performance in all other scenarios. These results suggest that the effectiveness of LN decompositions could be enhanced even more by incorporating covariance matrices that better reflect the dependency structures in X. Additionally, alternative hyperparameter settings for R^2-based priors might improve their performance, potentially closing the gap with the Horseshoe.

If we focus on the $\rho = 0.8, K = 750$ case in both Figures 13 and 14, we can see that there is a tradeoff between out-of-sample predictive performance and detecting signals. Here, the Dirichlet is not shrinking sufficiently (leading to worse out-of-sample predictions), while the LN might be overshrinking some signals (but ensures better out-of-sample predictions overall). We also show coverage properties using 95% marginal credibility intervals in Table 2 for the case where $\rho = 0$ and $R^2 = 0.6$. While frequentist properties are not guaranteed in a Bayesian framework, all models achieve the nominal 95% coverage independently of the value of K.

Specificity and sensitivity remain high for all values of K. However, for $K = 750$, models using the Dirichlet distribution (D2, DL) produce wider posterior credibility intervals, which negatively impacts sensitivity. Nevertheless, we still recommend a two-step procedure if selection is sought after.

We present ROC curves in Figure 15. The diagonal line shows the identity line. The LN models perform comparably to the Horseshoe and BP models, even surpassing

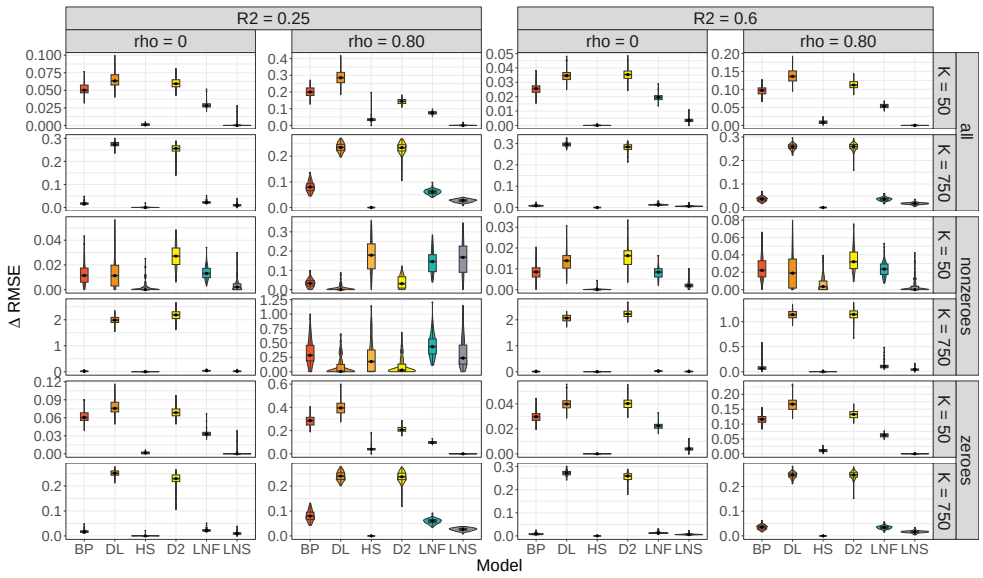


Figure 14: **Fixed coefficients setup** The true value of the signal is 3. ΔRMSE has been partitioned with respect to truly zero and nonzero coefficients.

the latter in some cases. In contrast, models using the Dirichlet distribution tend to underperform, likely due to their wider credibility intervals.

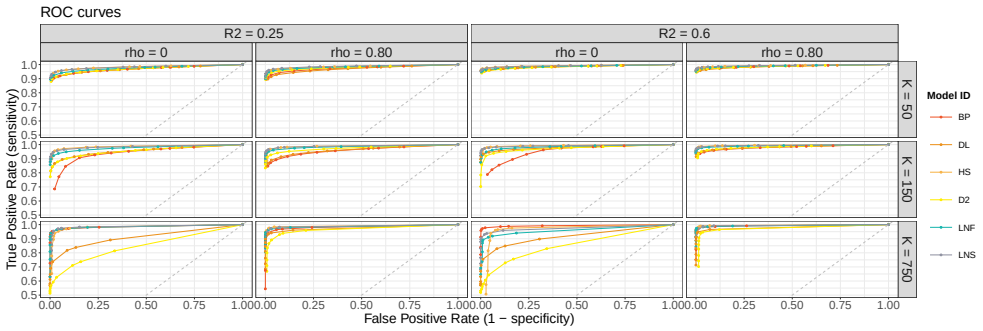


Figure 15: **Fixed coefficients setup:** ROC curves for different models under different simulation conditions. Points are obtained by varying the credibility level of marginal posterior credibility intervals. The grey dotted line depicts the identity line.

We also provide MCMC diagnostics. Figures 16, 17, and 18 illustrate the Effective Sample Size (ESS), Rhat, and total runtime results for the coefficient chains across the simulations (Vehtari et al., 2021). The Rhat diagnostic shows no major convergence

Table 2: **Fixed coefficients setup:** Coverage properties of 95% marginal posterior credible intervals for the scenario $\rho = 0$, $R^2 = 0.6$

K	Model ID	Coverage	Specificity	Sensitivity (Power)	Avg. CI Width
50	BP	0.956	0.982	0.986	0.369
	DL	0.971	0.976	0.997	0.387
	HS	0.987	0.996	0.997	0.351
	D2	0.960	0.983	0.986	0.398
	LNF	0.965	0.991	0.986	0.375
	LNS	0.970	0.996	0.987	0.358
150	BP	0.926	0.941	0.980	0.375
	DL	0.999	1.000	0.995	0.669
	HS	0.988	0.991	0.997	0.353
	D2	0.991	0.999	0.975	0.647
	LNF	0.988	0.997	0.985	0.404
	LNS	0.990	0.999	0.987	0.375
750	BP	0.998	1.000	0.989	0.375
	DL	0.989	1.000	0.222	2.585
	HS	0.971	0.972	0.998	0.341
	D2	0.987	1.000	0.078	2.409
	LNF	0.998	0.999	0.988	0.441
	LNS	0.998	1.000	0.989	0.388

issues for most models, with the exception of the Horseshoe and Beta Prime, which display the largest deviations from 1.0. ESS values are generally within acceptable ranges, indicating efficient sampling, although the Beta Prime demonstrates the lowest ESS in high-dimensional settings. In terms of runtime, LN models remain the fastest for low-complexity cases. As dimensionality increases, runtime slows across all models, though the increase in time is manageable and does not render any model impractical.

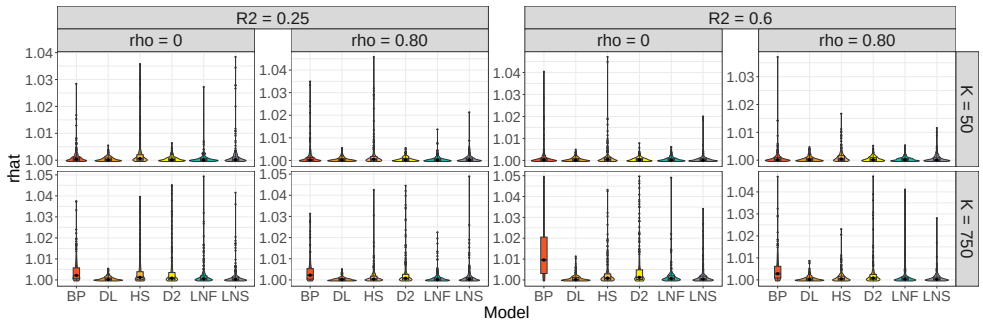


Figure 16: **Fixed coefficients setup:** Distribution of Rhat across different simulations.

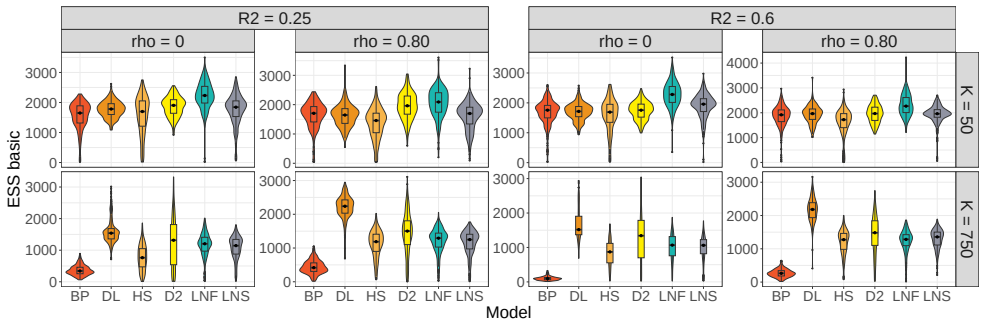


Figure 17: **Fixed coefficients setup:** Distribution of Effective Sample Size across different simulations.

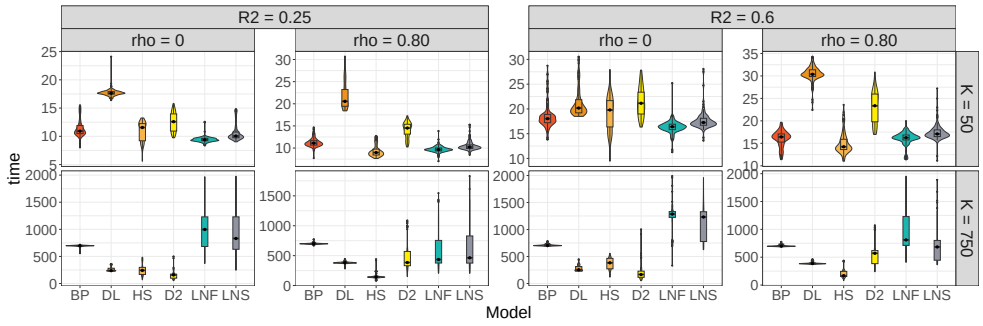


Figure 18: **Fixed coefficients setup:** Distribution of total sampling time for each model.

References

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711 – 1767.
URL https://doi.org/10.1214/23-EJS2136 3, 8

Aitchison, J. and Shen, S. M. (1980). "Logistic-Normal Distributions: Some Properties and Uses." *Biometrika*, 67(2): 261–272.
URL http://www.jstor.org/stable/2335470 1

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233.
URL https://doi.org/10.1080/02331888.2019.1662017 10

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490. PMID: 27019543.
URL https://doi.org/10.1080/01621459.2014.960967 8, 10

Carvalho, Polson, and Scott (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.
URL http://www.jstor.org/stable/25734098 10

Leydold, J. and Hörmann, W. (2011). "Generating generalized inverse Gaussian random variates by fast inversion." *Computational Statistics and Data Analysis*, 55(1): 213–217.
URL https://www.sciencedirect.com/science/article/pii/S0167947310002847 4

Neal, R. M. (2003). "Slice sampling." *The Annals of Statistics*, 31(3): 705–767. Publisher: Institute of Mathematical Statistics.
URL https://projecteuclid.org/journals/annals-of-statistics/volume-31/issue-3/Slice-sampling/10.1214/aos/1056562461.full 3, 4, 5

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155 – 2197.
URL https://doi.org/10.1214/20-EJS1711 12

Robert, C. (1991). "Generalized inverse normal distributions." *Statistics & Probability Letters*, 11(1): 37–41.
URL https://www.sciencedirect.com/science/article/pii/016771529190174P 3, 4

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer.
URL http://link.springer.com/10.1007/978-1-4757-4145-2 3, 4

Stuart, A. and Ord, K. (2009). *Kendall's Advanced Theory of Statistics: Volume 1: Distribution Theory*. Number vol. 1 ;vol. 1994 in Kendall's Advanced Theory of

Statistics. Wiley.
URL https://books.google.ch/books?id=tW18thQWJQIC 2

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667 – 718.
URL https://doi.org/10.1214/20-BA1221 13, 16

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13.
URL https://doi.org/10.1080/01621459.2020.1825449 3, 8, 10

Article III

Aguilar, J. E., Kohns, D., Vehtari, A., and Bürkner, P.-C. (2025). R2 priors for grouped variance decomposition in high-dimensional regression.

Preprint. Manuscript submitted for publication.

Available at https://arxiv.org/abs/2507.11833

R2 priors for Grouped Variance Decomposition in High-dimensional Regression

Javier Enrique Aguilar [*,‡] , David Kohns[†] , Aki Vehtari[†] and Paul-Christian Bürkner[*]

Abstract.

We introduce the Group-R2 decomposition prior, a hierarchical shrinkage prior that extends R2-based priors to structured regression settings with known groups of predictors. By decomposing the prior distribution of the coefficient of determination R^2 in two stages, first across groups, then within groups, the prior enables interpretable control over model complexity and sparsity. We derive theoretical properties of the prior, including marginal distributions of coefficients, tail behavior, and connections to effective model complexity. Through simulation studies, we evaluate the conditions under which grouping improves predictive performance and parameter recovery compared to priors that do not account for groups. Our results provide practical guidance for prior specification and highlight both the strengths and limitations of incorporating grouping into R2-based shrinkage priors.

Keywords: Prior specification, shrinkage priors, variance decomposition, regularization.

1 Introduction

In regression problems involving high-dimensional or structured predictors, covariates often have natural groupings that reflect scientific knowledge. Examples include genes in the same biological pathway (Li and Li, 2008; Sun et al., 2014), indicators of socioeconomic time-series (Mogliani and Simoni, 2021; Kohns and Potjagailo, 2025), or survey items that measure the same latent construct (Bentler and Bonett, 1980). In such cases, incorporating group structure into the prior distribution can improve estimation, interpretation, and predictive performance. However, many existing group-aware priors are motivated by marginal shrinkage properties and offer limited control over interpretable model summaries, such as the proportion of variance explained.

The coefficient of determination, R^2, is one such interpretable quantity which summarizes how well the model fits the data and is considered to be understood in many applied fields (Gelman et al., 2013). Shrinkage priors built around a prior distribution on R^2, such as the R2D2 prior and its extensions, allow users to express beliefs about model complexity and sparsity (Zhang et al., 2020; Aguilar and Bürkner, 2023; Yanchenko et al., 2025a; Kohns et al., 2025; Aguilar and Bürkner, 2025). As R2-type priors start from defining the prior distribution on R^2, they are predictively consistent priors. This

[*]TU Dortmund
[†]Aalto University
[‡]Corresponding author: javier.aguilarr@icloud.com url: https://jear2412.github.io

means that the prior predictive distribution does not substantially change when more model components are added. However, R2-based priors typically assume exchangeability across coefficients and do not consider structural information such as grouping. This can be suboptimal for parameter inference as well as prediction in structured regression settings where variance is expected to concentrate in specific subsets of predictors.

In this work, we introduce the Group-R2 prior, a class of continuous shrinkage priors that generalizes the R2-based framework to explicitly account for known grouping structures among predictors. Our approach builds on the idea of decomposing the total prior variance implied by an R^2 distribution, and extends it through a two-stage hierarchical allocation: first, the total variance is distributed across groups, and then within each group across individual coefficients. This leads to a prior that allows for targeted control over sparsity and signal dispersion at both the group and coefficient levels.

We show that the Group-R2 prior retains the desirable properties of global-local shrinkage priors (Polson and Scott, 2012), including heavy-tailed marginals, concentration at zero, and robustness to large signals. We additionally show how the shrinkage behavior of the prior depends on a small number of interpretable hyperparameters, and we propose strategies for setting them based on theoretical and shrinkage properties. These influence a measure of implicit model sizes, which the user can use to tune the prior. Many previous approaches to setting priors on known groups of coefficients have been similarly motivated from the global-local class of priors for exchangeable coefficients (Xu and Ghosh, 2015; Xu et al., 2016; Boss et al., 2023; Tew et al., 2025), or place discrete mass at zero for whole groups (Xu and Ghosh, 2015; Agarwal et al., 2025). Nonetheless, only the Group Inverse-Gamma Gamma (GIGG) prior of Boss et al. (2023) similarly belongs to the global-local family of priors where the prior consists of three scales: a global, common to all coefficients, a group-wise scale shared by all members of a group, and one individual to each coefficient. Unlike the Group-R2 prior, it is not built from a prior on R^2. The prior it induces on R^2 depends on its hyperparameters in a way that is not straightforward to control, changing as predictors are added, so the GIGG is not predictively consistent (Section 3.3).

We study the Group-R2 prior in simulations from two angles. First, to understand whether and when grouping information improves inference, we compare grouped and nongrouped versions of the R2 prior across scenarios that vary in signal strength, sparsity pattern and dimensionality. This isolates the effect of the additional decomposition and shows when grouping helps and when it does not. Second, we compare the Group-R2 prior with existing grouped priors. For the GIGG, we calibrate its hyperparameters so that it implies the same prior on R^2 as the Group-R2 prior, so that the two differ only in how the explained variance is decomposed (Section 5.5). We further compare against spike-and-slab versions of the Bayesian group lasso (Xu and Ghosh, 2015), the group horseshoe (Xu et al., 2016), grouped regression with adaptive shrinkage priors (GRASP) (Tew et al., 2025), and a grouped-informed g-prior (Agarwal et al., 2025) (Section 5.4). These comparisons show that the Group-R2 prior is competitive with existing grouped priors while giving users direct control over the prior on R^2, the quantity that most influences its performance, and that it shrinks real signals less than those priors that allow exact zeros.

2 Methods

In this section, we develop the Group-R2 prior, a hierarchical shrinkage prior that decomposes explained variance across and within predictor groups. We begin by detailing the construction of the prior in Section 2.1. We then present related priors in the literature that also account for grouped structures in Section 2.2. We introduce the concept of group-wise explained variance and related quantities in Section 2.3. We then derive key theoretical properties, including marginal and joint distributions, behavior near the origin and in the tails, and implications for shrinkage at both the group and coefficient levels in Section 2.4. We analyze the shrinkage behavior induced by the prior through the study of shrinkage factors and the effective number of non-zero coefficients in Section 2.5, and provide practical recommendations for hyperparameter specification in Section 2.6.

2.1 Group-R2 Decomposition priors

Our main interest lies in performing inference for the linear regression model in a situation in which the user already has knowledge about relevant groups of predictors. We will use the available group information to construct a multivariate continuous joint shrinkage prior that accounts for the grouping structures through the variance. Importantly, we do not assume the user has information about the dependency structure within the groups, only about which predictors belong to each mutually exclusive group (Müller and Quintana, 2004).

Let y denote the vector of n observations $(y_1, \ldots, y_n)'$, and let X be the design matrix of dimension $n \times p$. Suppose that we have prior knowledge of G groups. For each group $g = 1, \ldots, G$, let X_g be the matrix of predictors for that group, of dimension $n \times p_g$, where p_g denotes the column dimension of the gth group, and let the corresponding regression coefficients be $b_g = (b_{g1}, \ldots, b_{gp_g})'$. The overall design matrix can be written as $X = (X_1 \ldots X_G)$, with associated coefficient vector $b = (b_1' \ldots b_G')'$. The linear regression model under consideration is

$$y = b_0 J + \sum_{g=1}^{G} X_g b_g + \varepsilon, \tag{1}$$

where J is a vector of ones, $b_0 \in \mathbb{R}$ is the intercept and $\varepsilon = (\varepsilon_1 \ldots \varepsilon_n)'$ is such that $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$.

This model can be equivalently expressed observation-wise as

$$y_i = b_0 + \sum_{g=1}^{G} \sum_{j=1}^{p} x_{ij} b_j \mathbf{1}_{\{j \in \text{ group } g\}} + \varepsilon_i, \quad i = 1, ..., n, \tag{2}$$

where $\mathbf{1}$ is the indicator function. Priors that decompose the proportion of explained variance R^2 first define a global R^2 measure, which is then related to the variance of the linear predictor $x'b$ via a one-to-one transformation. Consequently, probabilistic

statements about R^2 directly translate to statements about the total variance. This total variance is subsequently decomposed to determine the scales of the coefficients b using an appropriate distribution, such as a normal or double exponential, to express the variability attributed to these coefficients.

We begin by defining the global R^2 measure as

$$R^2 := \text{Cor}^2(y, x'b) = \frac{\text{Var}(x'b)}{\text{Var}(x'b) + \sigma^2}. \tag{3}$$

In the following assume that $\mathbb{E}(x) = 0$ and $\text{Var}(x) = \Sigma_x$, where Σ_x has a diagonal of ones. We further assume the prior for b satisfies $\mathbb{E}(b) = 0$ and $\text{Var}(b) = \sigma^2\Lambda$, with Λ being a diagonal matrix whose entries are $\lambda_1^2, \ldots, \lambda_p^2$. Calculating the variance of the linear predictor yields $\text{Var}(x'b) = \sigma^2 \sum_{i=1}^{p} \lambda_i^2$. The quantity $\tau^2 := \sum_{i=1}^{p} \lambda_i^2$ is denoted as the total variance. This implies the following one-to-one relationship between R^2 and τ^2

$$R^2 = \frac{\tau^2}{\tau^2 + 1}. \tag{4}$$

Therefore probabilistic statements about the proportion of explained variance R^2 can be translated to the total variance τ^2 and vice versa. To incorporate group structures, we further decompose Λ into a block-diagonal matrix with entries $\Lambda_1, \ldots, \Lambda_G$, where $\Lambda_g = \text{diag}(\lambda_{g1}^2, \ldots, \lambda_{gp_g}^2)$. We can rewrite the total variance τ^2 as

$$\tau^2 = \sum_{g=1}^{G} \sum_{l=1}^{p_g} \lambda_{gl}^2 = \sum_{g=1}^{G} \tau_g^2, \tag{5}$$

where $\tau_g^2 = \sum_{l=1}^{p_g} \lambda_{gl}^2$ represents the group-wise total variance attributable to the gth group. We construct the Group-R2 prior through the following two-stage decomposition:

Stage I: Group variance decomposition. Decompose the total variance τ^2 across groups using groupwise proportions $\phi = (\phi_1, \ldots, \phi_G)$ such that $\phi_g \geq 0$ and $\sum_{g=1}^{G} \phi_g = 1$. The variance contribution of each group is then $\tau_g^2 = \phi_g \tau^2$.

Stage II: Within group decomposition. For each group g, decompose its total variance τ_g^2 into the variances of its coefficients. Denote the within-group proportions by $\varphi_g = (\varphi_{g1}, \ldots, \varphi_{gp_g})$, with $\varphi_{gl} \geq 0$ and $\sum_{l=1}^{p_g} \varphi_{gl} = 1$. The variances of the individual coefficients b_{gl} are then given by

$$\text{Var}(b_{gl} \mid \varphi_g, \phi, \tau^2, \sigma^2) = \lambda_{gl}^2 \sigma^2 = \varphi_{gl} \tau_g^2 \sigma^2 = \varphi_{gl} \phi_g \tau^2 \sigma^2, \quad l = 1, ..., p_g, \quad g = 1, \ldots, G. \tag{6}$$

To complete the specification of the joint prior, we assume that the R^2 prior follows a beta distribution with prior mean μ_{R^2} and precision ν_{R^2}. This facilitates an intuitive incorporation of domain knowledge into the R2 prior, as μ_{R^2} and ν_{R^2} can be directly communicated with users. From Equation 4, it follows that τ^2 is distributed according to a Beta Prime distribution with shape parameters a_1, a_2 (Bai and Ghosh, 2021), where

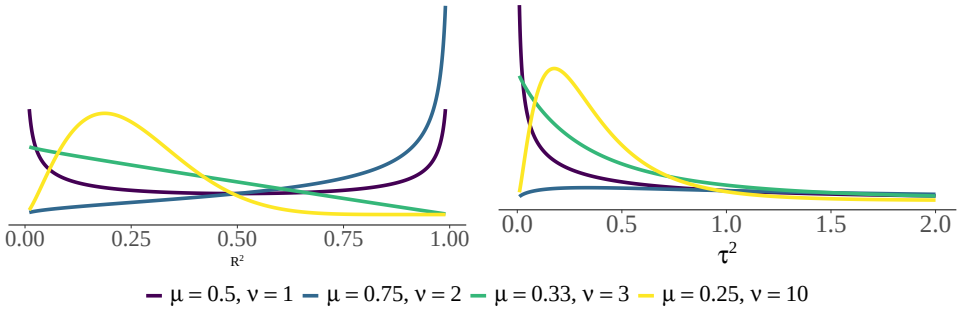


Figure 1: Beta and Beta Prime densities for several values of the prior mean μ_{R^2} and precision ν_{R^2}.

(a_1, a_2) denote the canonical parameters of the Beta distribution. Their relationship between the mean and precision is $\mu_{R^2} = a_1/(a_1 + a_2)$ and $\nu_{R^2} = a_1 + a_2$. We will make use of both parametrizations in the paper.

Figure 1 illustrates the flexibility of Beta distributions in encoding prior beliefs about R^2. It also shows the implied Beta Prime prior on τ^2. When the parameters are set to $(\mu_{R^2}, \nu_{R^2}) = (0.5, 1)$, the resulting prior on R^2 is bathtub-shaped, placing most of its mass near 0 and 1. This reflects a belief that the model is likely to explain either very little or nearly all of the variance—indicating an expectation of either predominantly noise or strong signal.

The proportions of group variance ϕ and within-group variance φ_g lie on the simplex. The canonical distribution for the simplex is the Dirichlet distribution, parameterized by the concentration vector, α (Bhattacharya et al., 2015; Lin, 2016). This vector, α, fully determines the mean and correlation structure of the Dirichlet distribution. A parsimonious choice within this family is the symmetric Dirichlet distribution, where $\alpha = (a_\pi, \ldots, a_\pi)$ with $a_\pi > 0$. We denote this distribution as $\phi \sim \text{Dirichlet}\, s(a_\pi)$, where $s(\cdot)$ emphasizes the symmetry of the concentration parameters. This choice greatly simplifies the number of hyperparameters to specify and facilitates the development of theoretical properties of the prior. [1]

To fully specify the joint hierarchical prior, we require a kernel for the coefficients b_{gl}. Consistent with the approaches of Aguilar and Bürkner (2023, 2025) and Kohns et al. (2025), we model b_{gl} as normally distributed with conditional variance $\text{Var}(b_{gl} \mid \lambda_{gl}^2, \sigma^2) = \sigma^2 \lambda_{gl}^2$.[2] The overall procedure to construct the Group-R2 prior is illustrated in Figure 2.1.

[1] For an alternative prior distribution for the simplex, see Aguilar and Bürkner (2025) who use the more flexible, logistic normal distribution. Due to difficulty of theoretical study, we leave further investigation with the logistic normal for future research.

[2] An alternative is presented in Zhang et al. (2020), who use a Laplace kernel to induce strong marginal concentration around zero. However, Schmidt and Makalic (2020) show that the asymptotic properties of this choice are equivalent to those obtained using a normal kernel.

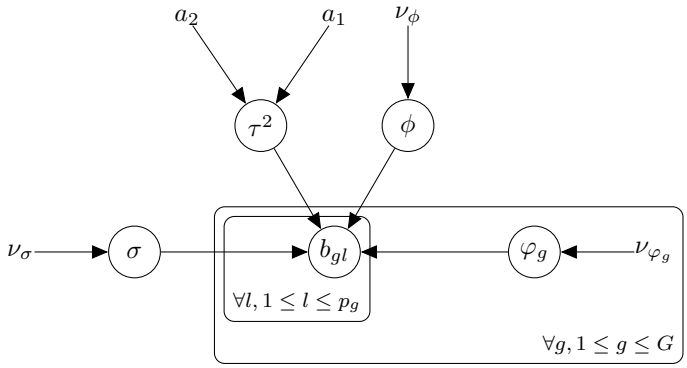


Figure 2: Directed acyclic graph (DAG) representing the prior structure of the Group-R2 decomposition prior. A global scale parameter τ^2, governed by hyperparameters a_1 and a_2, controls the total variance allocated across groups. The vector ϕ represents a group-level variance decomposition, drawn from a distribution over the simplex parameterized by ν_ϕ. For each group g, a within-group allocation vector φ_g is drawn from another simplex distribution with parameter ν_{φ_g}. Each coefficient b_{gl} inherits its prior variance through the product $\varphi_{gl}\phi_g\tau^2\sigma^2$. Plates indicate replication over groups ($g = 1, \ldots, G$) and coefficients within groups ($l = 1, \ldots, p_g$). This illustrates how variance is hierarchically allocated across and within groups.

We place a half Student-t prior on σ with η_σ degrees of freedom and scale parameter ν_σ, whose mean, when $\eta_\sigma > 1$ is proportional to ν_σ, and whose variance is proportional to ν_σ^2 when $\eta_\sigma > 2$ (Gelman, 2006; Goodrich et al., 2020). If a Gibbs sampler is preferred, an inverse gamma prior for σ^2 may be used as a computationally convenient alternative. The intercept b_0 is typically assigned either a normal prior or an improper uniform prior to ensure flexibility in modeling the overall mean.

The complete Group-R2 prior has the following form:

$$\begin{aligned} b_{gl} \mid \varphi_g, \phi, \tau^2, \sigma^2 &\sim \text{Normal}\left(0, \varphi_{gl}\phi_g\tau^2\sigma^2\right) \\ \phi \sim \text{Dirichlet}(\cdot),\ \varphi_g &\sim \text{Dirichlet}(\cdot), \\ \tau^2 \sim \text{BetaPrime}(\mu_{R^2}, \nu_{R^2}),\ \sigma &\sim \pi(\sigma). \end{aligned} \tag{7}$$

2.2 Related priors

The Group Inverse-Gamma Gamma (GIGG) prior, introduced by Boss et al. (2023), is a recent advancement in the class of continuous shrinkage priors that explicitly account for known grouping structures among covariates. Like the Group-R2 prior, this prior is designed for regression settings with block-correlated regressors, where effective shrinkage must balance signal detection and variance control across and within groups.

The GIGG prior extends the global-local shrinkage framework by introducing a third, group-specific layer of hierarchy. Each coefficient b_{gl} is modeled as conditionally normal

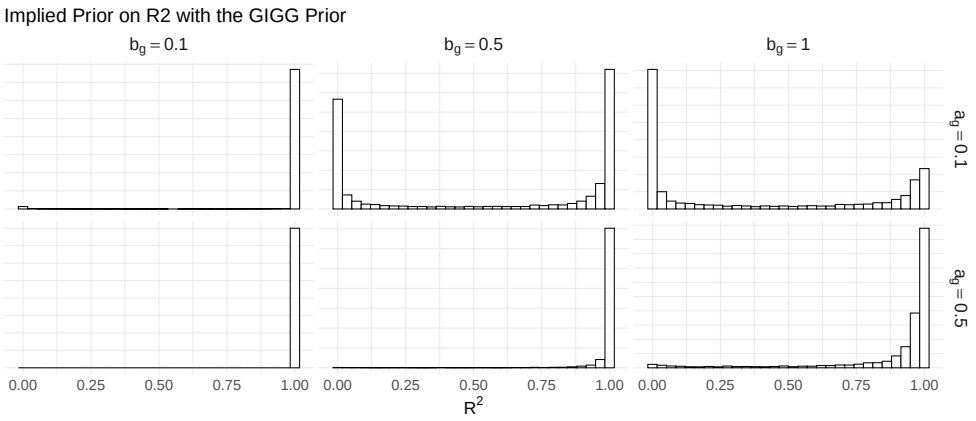


Figure 3: Implied prior distributions on R^2 under the GIGG prior for various combinations of a_g, b_g. We obtain the distributions via Monte Carlo and consider $G = 10$ groups, $p = 50$ covariates and we have fixed $\sigma = 1$. The hyperparameter a_g governs the strength of thresholding, since it controls the spike at zero in the marginal distribution of b_{gl}. In contrast, b_g regulates the degree of shrinkage dependence within each group, influencing how strongly individual coefficients borrow strength from one another. The induced R^2 distribution varies substantially with these parameters, often lacking direct interpretability or control.

with variance proportional to the product $\tau^2\gamma_g^2\lambda_{gl}^2$, where τ^2 is a global scale, γ_g^2 is a group-specific scale, and λ_{gl}^2 is a local (coefficient-specific) scale. These scales are assigned the priors $\gamma_g^2 \sim \text{Gamma}(a_g, 1)$ and $\lambda_{gl}^2 \sim \text{Inv-Gamma}(b_g, 1)$, where we parameterize the gamma distribution in terms of shape and rate. This hierarchical structure induces a beta prime distribution on the product $\gamma_g^2\lambda_{gl}^2$. We write the GIGG shapes as (a_g, b_g) throughout, reserving c_g for the within-group concentration of the Group-R2 prior.

Figure 3 illustrates the implied prior distribution on R^2 under the GIGG prior across various settings of (a_g, b_g). The results show that although the GIGG prior supports both group- and coefficient-level shrinkage, it does not offer explicit or intuitive control over the prior mass on R^2. The shape and spread of the induced R^2 distributions vary considerably with hyperparameter choices and can lead to undesired concentration away from interpretable target regions (e.g., low or moderate R^2 values). In contrast, the Group-R2 prior explicitly parametrizes R^2, allowing users to encode prior beliefs about explained variance both at a global and group-level more directly and transparently.

Other related priors include Bayesian Group Lasso and sparse group extensions that combine group-level and coefficient-level regularization through spike-and-slab components (Xu and Ghosh, 2015). The group horseshoe (Xu et al., 2016) and GRASP (Tew et al., 2025) consider continuous local scales, while group g-priors distribute mass to models themselves (Agarwal et al., 2025). In synthesis, these approaches differ in

whether coefficients can be exactly zero, how information is shared within a group, and how global regularization is specified.

While we were conducting this research, Yanchenko et al. (2025b) independently proposed a related group R2D2 prior that also employs a hierarchical decomposition of explained variance across groups and within groups. The two developments were carried out in parallel and were motivated by related ideas in R^2-based prior formulations.

2.3 Group-wise explained variances

In the following we introduce and characterize group-wise explained variances R^2_g, which quantify the share of total explained variance attributable to each predictor group. This helps interpret how the Group-R2 prior distributes variance across groups, complementing the global R^2 perspective. We derive properties of R^2_g and related quantities, such as the marginal distributions of group variances τ^2_g and coefficient specific variances λ^2_{gl}. These results will help us in our discussion on theory, shrinkage behavior, and hyperparameter specification shown thereafter.

Under the assumptions presented in Section 2.1, we define a group-specific explained variance metric R^2_g as:

$$R^2_g = \frac{\mathrm{Var}(x_g' b_g)}{\mathrm{Var}(x'b) + \sigma^2} = \frac{\tau^2_g}{\tau^2 + 1}, \quad g = 1, \ldots, G. \tag{8}$$

Equivalently,

$$R^2_g = \frac{\tau^2_g}{\tau^2 + 1} = \frac{\phi_g \tau^2}{\tau^2 + 1} = \phi_g R^2. \tag{9}$$

The quantity R^2_g can be interpreted as the proportion of the total variance explained by the predictors in the gth group. Proposition 1 characterizes the joint distribution of R^2_g and shows that different group contributions $R^2_g, R^2_h, h \neq g$ will exhibit negative dependencies conditional on R^2. Even though we could expect this since $\sum_{g=1}^{G} R^2_g = R^2$, mixing over random totals can change signs of the covariance (Kruijer et al., 2010). Proofs can be found in the Appendix.

Proposition 1. *(Joint distribution of R^2_g) If $\phi \sim \mathrm{Dirichlet}(\alpha)$ is independent of R^2, with $G \geq 2$ then the following holds.*

1. *The joint distribution of $(R^2_1/R^2, \ldots, R^2_G/R^2) = \phi \sim \mathrm{Dirichlet}(\alpha)$.*
2. *$R^2_g/R^2 = \phi_g \sim \mathrm{Beta}\left(\alpha_g, \sum_{j \neq g} \alpha_j\right)$.*
3. *The conditional distribution $\left(R^2_1, \ldots, R^2_G\right) \mid (R^2 = r^2) \sim r^2\,\mathrm{Dirichlet}(\alpha)$. Given $g \neq h$ then*

$$\mathrm{Cov}\left(R^2_g, R^2_h \mid R^2 = r^2\right) = (r^2)^2\,\mathrm{Cov}(\phi_g, \phi_h) = -(r^2)^2 \frac{\alpha_g \alpha_h}{\alpha_0^2(\alpha_0 + 1)}, \tag{10}$$

where $\alpha_0 = \sum_{i=1}^{G} \alpha_i$.

When $\phi \sim \text{Dirichlet}(\alpha)$, the distribution of $R_g^2 = \phi_g R^2$ corresponds to the product of two Beta-distributed random variables. A closed-form expression exists, however it is typically too cumbersome to be useful in practice, so we do not present it here. Its full derivation can be found in Coelho and Alberto (2021). However, the kth moment of R_g^2 can be found under a closed form and the set of them will completely determine their distribution (Schmüdgen, 2020):

$$\mathbb{E}[(R_g^2)^k] = \mathbb{E}[(\phi_g)^k]\mathbb{E}[(R^2)^k] = \left(\prod_{i=0}^{k-1} \frac{\alpha_g + i}{\alpha_0 + i}\right)\left(\prod_{j=0}^{k-1} \frac{a_1 + j}{a_1 + a_2 + j}\right), \quad k \in \mathbb{N}. \tag{11}$$

For instance if $\phi_g \sim \text{Beta}(\alpha_g, 1)$ and independently $R^2 \sim \text{Beta}(\alpha_g + 1/2, 1)$, then R_g^2 has the distribution of Z^2 with $Z \sim \text{Beta}(2\alpha_g, 2)$, as follows by comparing the moments above (Schmüdgen, 2020).

We characterize the marginal distributions of the total variance of the gth group τ_g^2 and the variances λ_{gl}^2 in Proposition 2. These results will be used in Section 2.4 to establish properties of our prior. Additionally, it will be useful for hyperparameter specification in Section 2.6.

Proposition 2. *(Marginals of τ_g^2 and λ_{gl}^2) Let $\tau^2 \sim \text{BetaPrime}(Ga_G, a_2)$, $\phi \sim \text{Dirichlet}\, s(a_G)$, and $\varphi_g \sim \text{Dirichlet}\, s(c_g)$ be independent, then:*

1. *The group-wise total variance $\tau_g^2 = \phi_g \tau^2 \sim \text{BetaPrime}(a_G, a_2)$ marginally and independently.*
2. *If $p_g c_g = a_G$, then the coefficient-specific variances $\lambda_{gl}^2 = \varphi_{gl}\tau_g^2 \sim \text{BetaPrime}(c_g, a_2)$.*

Proposition 2 demonstrates that aligning the concentration parameters across the two decomposition levels yields a closed-form expression for the marginal distribution of the coefficient specific variances λ_{gl}^2. Under this formulation, $c_g \leq a_G$, indicating that the within-group shrinkage will be stronger than the group-level shrinkage (Aguilar and Bürkner, 2025). Note that specifying the second-level concentration parameter c_g becomes immediate once the first-level parameter a_G is chosen. In other words, if we have an idea of the desired group-level shrinkage, the corresponding coefficient-level shrinkage follows directly. Conversely, if we instead have information about the within-group concentration parameters c_g, then a_G satisfies the restriction as long as all products $p_g c_g$ are equal. Alternatively, an asymmetric group Dirichlet with shapes $\alpha_g = p_g c_g$ and $a_1 = \sum_g \alpha_g$ yields the analogous marginal identities.

Proposition 3 (Flat Dirichlets)**.** *If $\phi \sim \text{Dirichlet}(a_1^*, \ldots, a_G^*)$ and independently $\varphi_g \sim \text{Dirichlet}(c_{g1}, \ldots, c_{gp_g})$, with $a_g^* = \sum_l c_{gl}$, then the concatenated weights $(\phi_g \varphi_{gl})_{g,l}$ have a single Dirichlet distribution with coordinate shapes $(c_{gl})_{g,l}$.*

Therefore setting $a_G = p_g c_g$ produces a flat Dirichlet allocation vector. If the groups have the same size with a common c_g, the prior treats the partitions equally. The general

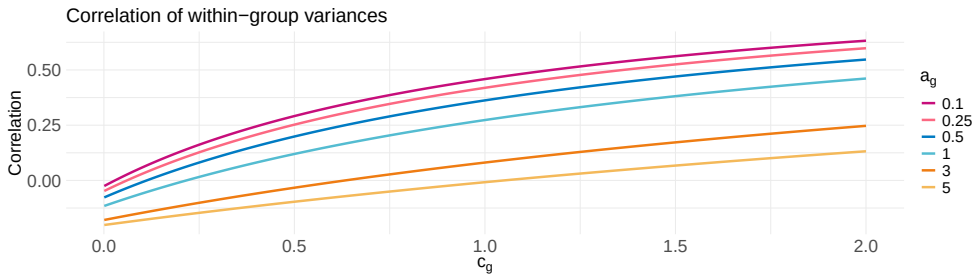


Figure 4: Correlation between within-group variance allocations $\phi_g\varphi_{gj}$ and $\phi_g\varphi_{gk}$ $(j \neq k)$ as a function of the within-group concentration c_g, for different values of the group-level concentration a_G. Positive correlations correspond to more distributed signal within a group, whereas negative correlations correspond to more concentrated signal. Results are shown for $G = 25$ groups with $p_g = 5$ covariates per group.

construction can express additional group dependencies by choosing the two allocation concentrations separately, which we show in the following.

Proposition 4 characterizes the dependence structure induced by the Group-R2 prior on the marginal variance proportions $\phi_g\varphi_{gj}$. In particular, it quantifies both the covariance and correlation of their logarithms, $\log(\phi_g\varphi_{gj})$ and $\log(\phi_g\varphi_{gk})$, for $j \neq k$. These log-scale quantities have been highlighted by Schmidt and Makalic (2020) and Tew et al. (2025) as useful to understand the concentration and tail behavior of shrinkage priors. We extend the results of Tew et al. (2025) to the case in which dependencies are considered for the within-group scales, allowing us to capture richer covariance structures than priors with independent scales.

Proposition 4 (Within-group variance dependencies)**.** *Assume that ϕ and φ_g are independent, ϕ is symmetric, and $p_g \geq 2$. Let ϕ_g denote the proportion of total variance allocated to group g, and let $\varphi_g = (\varphi_{g1}, \ldots, \varphi_{gp_g})$ denote the within-group allocation proportions for the p_g coefficients in group g. Then:*

1. *The correlation between the log-scaled variances of two coefficients within the same group satisfies*

$$\operatorname{Corr}\left(\log(\phi_g\varphi_{gj}),\ \log(\phi_g\varphi_{gk})\right) = \frac{\operatorname{Var}(\log\phi_g) + \operatorname{Cov}(\log\varphi_{gj}, \log\varphi_{gk})}{\operatorname{Var}(\log\phi_g) + \operatorname{Var}(\log\varphi_{g\cdot})}. \tag{12}$$

2. *Suppose $\tau^2 \sim \operatorname{BetaPrime}(Ga_G, a_2)$, $\phi \sim \operatorname{Dirichlet} s(a_G)$, and $\varphi_g \sim \operatorname{Dirichlet} s(c_g)$ independently. Then:*

   - *The correlation of log-variances is given by*

$$\operatorname{Corr}\left(\log(\phi_g\varphi_{gj}),\ \log(\phi_g\varphi_{gk})\right) = \frac{\psi_1(a_G) - \psi_1(Ga_G) - \psi_1(p_gc_g)}{\psi_1(a_G) - \psi_1(Ga_G) + \psi_1(c_g) - \psi_1(p_gc_g)}, \tag{13}$$

where $\psi_1(x) = d^2 \log \Gamma(x)/dx^2$ *is the trigamma function and* $\Gamma(x)$ *is the Gamma function (Olver et al., 2010b).*

- *The covariance of the raw allocation proportions is*

$$\operatorname{Cov}\left(\phi_g \varphi_{gj},\ \phi_g \varphi_{gk}\right) = \frac{a_G(a_G+1)}{G a_G(G a_G+1)} \cdot \frac{c_g}{p_g(c_g p_g+1)} - \frac{1}{G^2 p_g^2}. \tag{14}$$

Our framework introduces structured dependencies among the variances assigned to coefficients by using Dirichlet priors for both ϕ and φ_g. The second decomposition stage, parameterized by c_g, plays a key role in controlling these within-group correlations. As shown in Figure 4, the correlation between $\phi_g \varphi_{gj}$ and $\phi_g \varphi_{gk}$ increases with c_g, with the speed of this increase controlled by the group-level parameter a_G. Notably, for small values of c_g, the correlation can be negative, a behavior that cannot arise under priors like GIGG or GRASP (Tew et al., 2025), where group-specific scales are modeled independently.

2.4 Marginal and joint distributions

In this section, we explore the marginal distribution of the coefficients b_{gl} and the joint distribution of the coefficients within a group when τ_g^2 is specified. We demonstrate that the marginal coefficients exhibit polynomial tails and high concentration near the origin. Furthermore, we show that the joint distribution of the coefficients within each group displays a pronounced spike at zero. We also show that our prior can represent the horseshoe prior as a special case (Carvalho et al., 2010), hence expecting it to inherit the desirable properties from the latter.

For simplicity, we set $\sigma^2 = 1$ in the following derivations, although we note that all results are conditional on σ^2, consistent with common practice in the analysis of marginal distributions in shrinkage priors (Bhattacharya et al., 2015; Pas, 2021; Bai and Ghosh, 2021). We denote the Beta and Gamma functions by $\mathrm{B}(\cdot,\cdot)$ and $\Gamma(\cdot)$, respectively, and use $U(\eta, \nu, z)$ to represent the confluent hypergeometric function of the second kind (Jeffrey et al., 2007; Olver et al., 2010b). In the following we assume that $a_G > 0$, $R^2 \sim \mathrm{Beta}(G a_G, a_2)$, $\phi \sim \mathrm{Dirichlet}\, s(a_G)$, and $\varphi_g \sim \mathrm{Dirichlet}\, s(c_g)$ with $c_g = a_G/p_g$.

Proposition 5 (Marginal distributions of b_{gl})**.** *The marginal prior density of* b_{gl} *for any* $g = 1, \ldots, G$ *and* $l = 1, \ldots, p_g$ *is*

$$p(b_{gl}) = \frac{1}{\sqrt{2\pi}\ \mathrm{B}(c_g, a_2)} \Gamma(\eta) U(\eta, \nu_g, z_{gl}), \tag{15}$$

where $\eta = a_2 + 1/2$, $\nu_g = 3/2 - c_g$, *and* $z_{gl} = |b_{gl}|^2/2$.

Proposition 6 (Marginal distributions of b_{gl} near the origin)**.** *As* $|b_{gl}| \to 0$, *with* $0 < c_g \leq 1/2$ *and* $a_2 > 0$ *the marginal prior densities are unbounded and exhibit a singularity at zero. Specifically, they satisfy*

$$p(b_{gl}) \sim \begin{cases} c_1 \left|b_{gl}\right|^{2c_g - 1} + \mathcal{O}\left(\left|b_{gl}\right|^{2c_g+1}\right), & c_g < \frac{1}{2}, \\ -c_2 \ln(b_{gl}^2) + \mathcal{O}\left(b_{gl}^2 \ln(b_{gl}^2)\right), & c_g = \frac{1}{2}, \end{cases} \tag{16}$$

where $c_1, c_2 > 0$. Conversely, when $c_g > \frac{1}{2}$, the marginal prior densities are bounded and continuous at zero. Furthermore, they are differentiable at $b_{gl} = 0$ for all $c_g > 1$.

Proposition 6 shows that the interplay between the two levels of decomposition directly governs the behavior of the marginal priors near the origin, which in turn determines the amount of shrinkage imposed on the coefficients. Specifically, since the proposition requires $c_g \leq a_G$, it suggests that the within-group shrinkage is stronger than the group-level shrinkage. This ensures sufficient prior mass near zero, a crucial property for shrinkage priors to effectively detect and suppress irrelevant signals (Pas et al., 2014, 2017; Pas, 2021).

As c_g decreases from 1/2 to zero, the prior shifts an increasing amount of mass toward the origin, resulting in posterior distributions that provide sparse estimates of the coefficient vector b. Conversely, when c_g exceeds 1/2, the marginal priors become bounded at zero, favoring less sparse coefficient estimates. Thus, the parameter c_g offers a flexible mechanism for translating prior beliefs about sparsity (or the lack thereof) into the prior model, controlling the degree of shrinkage applied to the coefficients.

The tails of the prior play a crucial role in determining whether large signals can be detected and how strongly these signals are pulled back toward zero (Armagan et al., 2013; Pas et al., 2016; Pas, 2021). Ideally, priors with heavy tails are favored because they protect substantial signals from being overly shrunk (Carvalho et al., 2010; Piironen and Vehtari, 2017). This results in priors of bounded influence, meaning that truly large coefficients remain relatively unshrunken by the prior, i.e., if $X = I$, then $\mathbb{E}[b_i \mid y_i] \approx y_i$ when y_i is greater than a threshold. This phenomenon, also known as tail robustness (Carvalho et al., 2010; Pas, 2021), is essential in sparse settings to selectively shrink only the smallest coefficients while leaving larger, genuine signals largely untouched. We show in Proposition 7 that the Group-R2 prior attains heavier tails than the Cauchy distribution and then establish in Proposition 8 that it can represent the horseshoe prior and is therefore of bounded influence.

Proposition 7 (Tails of the marginal prior)**.** *As $|b_{gl}| \to \infty$, the marginal prior density of b_{gl} satisfies*

$$p(b_{gl}) \sim \mathcal{O}\left(\frac{1}{|b_{gl}|^{2a_2+1}}\right). \tag{17}$$

In particular, when $0 < a_2 < 1/2$, the marginal priors have heavier tails than the Cauchy distribution and have the same polynomial order when $a_2 = 1/2$.

Proposition 8 (Horseshoe type behavior)**.** *If $c_g = a_2 = \frac{1}{2}$, then $\lambda_{gl} \sim Cauchy^{+}(0, 1)$. The marginal distribution of b_{gl} is the horseshoe marginal with global scale one. This implies that the Group-R2 prior has bounded influence.*

The following proposition shows that we can still achieve poles at zero and Cauchy-order tails even without the coupling. However, this does not lead to exact horseshoe marginals in general.

Proposition 9 (Marginal behavior without coupling)**.** *Assume $a_1 = Ga_G$, $a_G, a_2, c_g > 0$, and fix $\sigma > 0$. For $p_g > 1$, the density of $b_{gl} \mid \sigma$ is finite at zero if and only if both*

$a_G > 1/2$ and $c_g > 1/2$, otherwise it diverges as $|b_{gl}| \to 0$. Moreover,

$$p(b_{gl}) \sim \frac{2^{a_2}\Gamma(a_2 + 1/2)\mathbb{E}(\varphi_{gl}^{a_2})}{\sqrt{\pi}B(a_G, a_2)}|b_{gl}|^{-2a_2-1}. \tag{18}$$

For a singleton group, $\varphi_{g1} = 1$ and finiteness at the origin requires only $a_G > 1/2$.

We can find the joint distribution of the coefficients within a group by integrating over the group variance parameter $\tau_g^2 = \phi_g\tau^2$. This is a similar approach to the idea presented by Boss et al. (2023), where integrating out variance parameters leads to conditional densities for grouped coefficients.

Proposition 10 (Joint distribution of group coefficients)**.** *Assume that $\tau_g^2 \sim \text{BetaPrime}(a_G, a_2)$ independently of φ_g. Conditional on τ_g^2, φ_g, and σ, let*

$$b_g \mid \tau_g^2, \varphi_g \sim \text{Normal}_{p_g}\left(0, \tau_g^2\,\text{diag}(\varphi_{g1}, \ldots, \varphi_{gp_g})\right). \tag{19}$$

Then the joint density of the coefficients in group g, conditional on φ_g, is

$$p(b_g \mid \varphi_g) = \frac{(2\pi)^{-p_g/2}}{B(a_G, a_2)}\left(\prod_l \varphi_{gl}\right)^{-1/2}\Gamma(\eta_g)U(\eta_g, \nu_g, z_g), \tag{20}$$

where $\eta_g = a_2 + p_g/2, \nu_g = 1 + p_g/2 - a_G, z_g = 1/2\sum_l b_{gl}^2/\varphi_{gl}$. Moreover, if φ_g is in the interior of the simplex, the joint density diverges at the group origin whenever $a_G < p_g/2$. When $a_G = p_g/2$, the divergence is logarithmic, and is finite when $a_G > p_g/2$.

The divergence of the joint density at zero for $c_g \leq 1/2$ indicates that the prior allocates substantial mass near zero, resulting in an infinite spike at the origin. This implies that we are able to shrink entire groups of coefficients if they are deemed as noise. Notably, when $p_g = 1$, the group consists of a single coefficient, and we recover the marginal prior distributions presented in Proposition 5. Equation 20 does not exhibit a traditional dependence structure, at least not one that is immediately apparent from its analytical form. To properly analyze the dependency structures implied by this joint prior, one should move beyond conventional multivariate dependency metrics and consider alternative representations, such as the implied copula.

2.5 Shrinkage properties

In this section, we examine how the prior influences shrinkage at both the coefficient and group levels. We first analyze the distribution of group shrinkage factors κ_g and marginal shrinkage factors κ_{gl}, which determine how much each group is jointly regularized and how much each coefficient is pulled toward zero under the normal means model. We then study the effective number of nonzero coefficients, m_{eff}, as a global measure of sparsity. These results show how hyperparameter choices translate into different regularization patterns.

Shrinkage Factors

As is common in the shrinkage literature, we can characterize the effect of the prior through the properties of the posterior distribution. In particular, analyzing the posterior mean and variance of each coefficient b_{gl} provides insight into how the prior induces shrinkage. Assume the normal means setting, where $X = I$ and $p = n$. Then the posterior distribution of b_{gl} given y_{gl} and its prior variance $\lambda_{gl}^2 = \varphi_{gl}\phi_g\tau^2\sigma^2$ is:

$$b_{gl} \mid y_{gl}, \lambda_{gl}^2 \sim \text{Normal}\left(\frac{\lambda_{gl}^2}{1+\lambda_{gl}^2} y_{gl}, \frac{\lambda_{gl}^2}{1+\lambda_{gl}^2}\sigma^2\right). \tag{21}$$

In this context, the maximum likelihood estimate (MLE) of b_{gl} is simply y_{gl} (Castillo and Vaart, 2012). The quantity $\kappa_{gl} = \frac{1}{1+\lambda_{gl}^2} \in (0,1)$ measures the degree of shrinkage imposed on the MLE. When the variance λ_{gl}^2 is large, little shrinkage is applied and $\kappa_{gl} \to 0$, whereas small variances lead to stronger shrinkage with $\kappa_{gl} \to 1$. The posterior mean of b_{gl} is directly related to κ_{gl} via

$$\mathbb{E}\left(b_{gl} \mid y_{gl}\right) = \left(1 - \mathbb{E}(\kappa_{gl} \mid y_{gl})\right) y_{gl}. \tag{22}$$

Hence, studying the distribution of κ_{gl} is key to understanding when the posterior of b_{gl} will be centered near zero or not. This connection makes κ_{gl} a useful diagnostic for assessing effective shrinkage, especially in hierarchical or group-structured priors (Castillo et al., 2015; Pas et al., 2014, 2016; Zhang and Bondell, 2018; Bai and Ghosh, 2019).

Let $\kappa_g = (\kappa_{g1}, \ldots, \kappa_{gp_g})'$ denote the vector of shrinkage factors for the gth group. Assume $\sigma^2 = 1$ and $R^2 \sim \text{Beta}(Ga_G, a_2)$ such that $\tau_g^2 = \phi_g\tau^2 \sim \text{BetaPrime}(a_G, a_2)$. We use the parametrization $\lambda_{gl}^2 = \varphi_{gl}\tau_g^2$, with $\varphi_g = (\varphi_{g1}, \ldots, \varphi_{gp_g})$ distributed as a symmetric Dirichlet with parameter c_g.

Note that the local scale of each coefficient is bounded above by the group scale, since $\varphi_{gl} \in (0,1)$ and $\sum_{l=1}^{p_g} \varphi_{gl} = 1$ implies $\lambda_{gl}^2 \leq \tau_g^2$. As a consequence, even when τ_g^2 is moderately large, some λ_{gl}^2 may remain small due to the Dirichlet structure, leading to persistent shrinkage for certain coefficients. In particular, if τ_g^2 is small, all λ_{gl}^2 will be small as well, and the group will experience strong shrinkage uniformly across its elements.

Conditional on $\tau_g^2 = t$, the vector κ_g lies on

$$\mathcal{K}(t) = \left\{\kappa_g \in (0,1)^{p_g} \mid \sum_{l=1}^{p_g} \tfrac{1-\kappa_{gl}}{\kappa_{gl}} = t\right\}. \tag{23}$$

The shrinkage vector κ_g is the pushforward distribution of Dirichlet vector φ_g and has no density with respect to the Lebesgue measure on the cube $(0,1)^{p_g}$. For a singleton group $\kappa_{gl} = (1+t)^{-1}$ deterministically. For $p_g \geq 2$ the moments are given by

$$\mathbb{E}\left(\kappa_{gl}^m \mid t\right) = {}_2F_1\left(m, c_g, c_g p_g, -t\right), \quad m \geq 0, \tag{24}$$

where ${}_2F_1(\alpha, \beta, \gamma, z)$ denotes the Gauss hypergeometric function (Jeffrey et al., 2007).

Even though we do not have access to the joint density of κ_g, we can still study the dependence structure of its elements. The following proposition shows that conditional competition among the weights coexists with dependence induced by the common group scale τ_g^2. Increasing the shrinkage of one component forces other components to be less shrunk.

Proposition 11 (Dependence of shrinkage factors)**.** *For $p_g \geq 2$, $c_g > 0$ and $t > 0$, the shrinkage factors satisfy $\mathrm{Cov}(\kappa_{gj}, \kappa_{gk} \mid \tau_g^2 = t) < 0$. If the group scale is integrated out, the sign of the covariance might change; when the group scale is integrated out and $a_1 = Ga_G$ and $p_g c_g = a_G$ the covariance is positive.*

We now derive the unconditional joint density of the shrinkage factors $\kappa_g = (\kappa_{g1}, \ldots, \kappa_{gp_g})'$. Its relevance comes from the fact that it describes the prior probability of different shrinkage configurations within a group such as complete group shrinkage, concentrated signal, and distributed signal. To do this we first derive the joint distribution of the within group scales λ_{gl}^2 and use the change of variables formula with $\lambda_{gl}^2 = \frac{1-\kappa_{gl}}{\kappa_{gl}}$. In the following let $a_1 = Ga_G$, $u_{gl} = \lambda_{gl}^2$, u_g be the vector collecting u_{gl}, and $s_g(\kappa_g) = \sum_{l=1}^{p_g} \frac{1-\kappa_{gl}}{\kappa_{gl}}$. Using the fact that $\tau_g^2 = \sum_{l=1}^{p_g} u_{gl} = s_g(\kappa_g)$ and $\varphi_{gl} = \frac{u_{gl}}{\sum_{r=1}^{p_g} u_{gr}}$ and $\tau_g^2 \sim \mathrm{BetaPrime}(a_G, a_2), \phi_g \sim \mathrm{Dirichlet}\, s(c_g)$ independently we arrive at the joint distribution

$$p(u_g) = \frac{\Gamma(p_g c_g)}{\Gamma(c_g)^{p_g} B(a_G, a_2)} \left(\sum_l u_{gl} \right)^{a_G - p_g c_g} \left(1 + \sum_l u_{gl} \right)^{-a_G - a_2} \prod_l u_{gl}^{c_g - 1}. \tag{25}$$

Applying the component-wise transformation $u_{gl} = \frac{1-\kappa_{gl}}{\kappa_{gl}}$ we obtain the joint density of κ_g on $(0,1)^{p_g}$:

$$p(\kappa_g) = \frac{\Gamma(p_g c_g)}{\Gamma(c_g)^{p_g} B(a_G, a_2)} s_g(\kappa_g)^{a_G - p_g c_g} \{1 + s_g(\kappa_g)\}^{-a_G - a_2} \prod_l (1 - \kappa_{gl})^{c_g - 1} \kappa_{gl}^{-c_g - 1}. \tag{26}$$

The dependence structure of the joint density is expressed through the terms containing $s_g(\cdot)$. Finally, we also show the distribution of the marginal shrinkage factor κ_{gl}. Marginally, Proposition 2 implies that $\lambda_{gl}^2 \sim \mathrm{BetaPrime}(c_g, a_2)$. Setting $\kappa_{gl} = \frac{1}{1+\lambda_{gl}^2}$ we have

$$p(\kappa_{gl}) = \frac{1}{B(a_2, c_g)} \kappa_{gl}^{a_2 - 1} (1 - \kappa_{gl})^{c_g - 1}, \qquad 0 < \kappa_{gl} < 1. \tag{27}$$

Hence $\kappa_{gl} \sim \mathrm{Beta}(a_2, c_g)$. This representation gives a direct interpretation of the two hyperparameters. Values of κ_{gl} close to one correspond to strong shrinkage, since λ_{gl}^2 is close to zero. The behavior near $\kappa_{gl} = 1$ is governed by c_g. Thus, $c_g < 1$ places high prior density near complete shrinkage. Values of κ_{gl} close to zero correspond to weak shrinkage and large coefficient-specific variances. The behavior near $\kappa_{gl} = 0$ is governed by a_2. Hence, $a_2 < 1$ places high prior density near no shrinkage. In particular, when

$a_2 = c_g = 1/2$, the shrinkage factor has the classical horseshoe shape $\kappa_{gl} \sim \text{Beta}(1/2, 1/2)$ (Carvalho et al., 2010).

Effective number of non-zero coefficients

A global summary of the shrinkage induced by the prior is provided by the effective number of nonzero coefficients, m_{eff}, defined as (Piironen and Vehtari, 2017)

$$m_{\text{eff}} = \sum_{i=1}^{p} (1 - \kappa_i), \tag{28}$$

where κ_i denotes the shrinkage factor associated with the ith coefficient b_i, where by construction $m_{\text{eff}} \leq p$. This quantity captures how the combination of global and local scales, encoded through κ_i, reduces model complexity.

Visualizing the prior distribution of m_{eff} is a practical tool for understanding how the choice of hyperparameters controls shrinkage. If the user has prior knowledge or expectations about the number of relevant coefficients, they can simulate the distribution of m_{eff} under different hyperparameter settings and select the configuration that aligns with their prior beliefs (Piironen and Vehtari, 2017; Aguilar and Bürkner, 2025). This approach is particularly useful in high-dimensional applications, where direct specification of priors on individual coefficients may be unintuitive or impractical. We can also define a group-wise effective number of nonzero coefficients as

$$m_{\text{eff},g} = \sum_{l=1}^{p_g} (1 - \kappa_{gl}), \tag{29}$$

so that the total effective number of nonzero coefficients decomposes as

$$m_{\text{eff}} = \sum_{g=1}^{G} m_{\text{eff},g}. \tag{30}$$

Under strong global shrinkage, we expect m_{eff} to concentrate near zero, while under weak global shrinkage, it may approach p. Similarly, for each group g, strong shrinkage at the group level (e.g., small values of τ_g^2) will result in $m_{\text{eff},g}$ concentrating near zero, indicating that most coefficients in the group are shrunk strongly toward zero. In contrast, when τ_g^2 is large, we expect $m_{\text{eff},g}$ to approach p_g, reflecting a weakly regularized group with more non-negligible coefficients. Figure 5 displays the prior predictive distribution of $m_{\text{eff},g}$, under varying values of a_G and c_g. Recall that a_G is the parameter of the symmetric Dirichlet distribution used in the first-stage decomposition (across groups), while c_g controls the second-stage allocation of variance within each group. Values of (a_G, c_g) below 1 push mass toward the edges of the simplexes, inducing sparsity in b_{gl}. The value of a_G also controls the prior mean of R^2, via $R^2 \sim \text{Beta}(Ga_G, a_2)$.

As a_G increases, the expected prior mean does so as well, from approximately 0.8 when $a_G = 0.1$ to around 0.99 when $a_G = 5$. When $a_G < 1$, the Dirichlet distribution

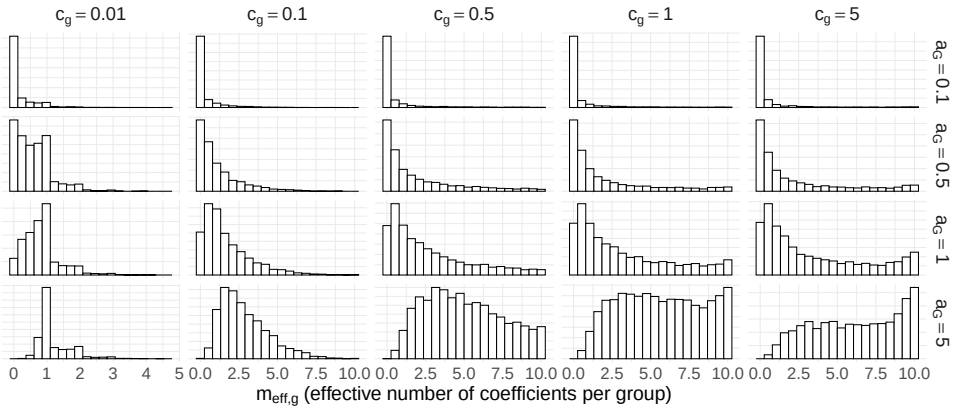


Figure 5: Prior predictive distributions of the group-wise effective number of non-zero coefficients ($m_{\text{eff},g}$) for different combinations of a_G and c_g. Here, a_G is the concentration parameter of the group-level Dirichlet decomposition (controlling between-group shrinkage) and c_g is the within-group concentration (controlling within-group shrinkage). Rows vary a_G and columns vary c_g. Results are shown for $G = 10$ groups with 10 covariates per group.

concentrates near the corners of the simplex, allocating most variance to a few groups. In contrast, for $a_G \geq 1$, the distribution tends to allocate variance more evenly across groups.

Given a fixed a_G, the total variance, τ_g^2, assigned to a group is limited by the Dirichlet partition. This constraint is evident in Figure 5: when a_G is small, stronger shrinkage occurs within groups regardless of c_g. Within-group sparsity is influenced by c_g. In line with this, we see that smaller values of c_g promote sparsity concentrating mass toward lower $m_{\text{eff},g}$, while larger values of c_g shift the prior mass toward less sparse configurations. These prior predictive distributions will naturally depend on the tail behavior of the prior, influenced largely by a_2. In the Appendix, we generate the figure for $a_2 = 0.1$.

2.6 Hyperparameter specification

Here we summarize based on the theoretical discussion above, some ways in which the prior hyperparameters (a_1, a_2, a_G, c_g) can be selected. Recall that $R^2 \sim \text{Beta}(a_1, a_2)$, $\phi \sim \text{Dirichlet}\, s(a_G)$ and $\varphi_g \sim \text{Dirichlet}\, s(c_g)$.

Coherent with the theoretical discussion in Section 2.4, one may significantly reduce the decision space by adopting $R^2 \sim \text{Beta}(Ga_G, a_2)$ and $c_g = a_G/p_g$. This ensures that the marginal prior distribution, $p(\lambda_{gl}^2)$, has a closed form Beta prime distribution (see Proposition 2) and gives an automatic way to choose c_g. This coupling allows the user

to tune group-level shrinkage via a_G, while inducing coherent within-group behavior. The user may then set a_G and a_2 based on prior information about R^2 from past studies or expert knowledge in the form of location and scale, using $\mu_{R^2} = a_1/(a_1 + a_2)$ and $\nu_{R^2} = a_1 + a_2$.

We have established that by setting $a_2 \in (0, 1/2)$, the marginal prior, $p(b_{gl})$, has heavier than Cauchy tails (Proposition 7). This makes the prior behave more and more like a variable selection type prior, the smaller a_2 is set within this range. When $a_2 = c_g = 1/2$, the prior achieves bounded influence (Proposition 8) such that we recommend setting $a_2 = 1/2$ as a flexible default choice.

If knowledge of the signal type within group is present, then a_G along with c_g can be set according to the desired joint shrinkage behavior. Figure 5 shows that larger a_G should be chosen for less shrinkage, or when many significant signals are expected in a given group (distributed signal setting) and low a_G, to exert stronger shrinkage. This is beneficial for when only few signals are expected (concentrated signal setting), or many small signals. For a sensible default choice of c_g, Figure 4 suggests that $c_g = 1/2$ allows a_G to control a wide range of correlation in variances. Hence, for a_G relatively small, signals are shrunk jointly very heavily to zero in a given group, while a_G relatively large, encourages relatively little shrinkage with little or negative correlation in variances. When one wants to express general lack of knowledge, a uniform R^2 prior $a_1 = a_2 = 1$ and uniform priors over the simplexes $a_G = c_g = 1$ may be chosen. This, however, does not lead to uniformity in the shrinkage coefficient space (Figure 5).

Boss et al. (2023) propose using Marginal Maximum Likelihood Estimation (MMLE), an empirical Bayes approach, to automatically estimate hyperparameters. They implement MMLE iteratively within their Gibbs sampler for posterior inference. The benefit of automatic selection is, however, offset by the additional computational complexity it introduces and is contrary to the focus of this paper to set hyperparameters based on theoretical principles or understanding of the model.

3 Experiments

We have implemented the models with the probabilistic programming language and framework Stan (Carpenter et al., 2017; Stan Development Team, 2024), which employs an adaptive Hamiltonian Monte Carlo (HMC) sampler known as the No-U-Turn Sampler (NUTS) (Neal, 2011; Brooks et al., 2011; Hoffman and Gelman, 2014) to sample draws from posterior distributions. The associated data and code are available at `https://github.com/jear2412/GroupR2priors`.

3.1 Evaluation Metrics

Our goal is to evaluate whether incorporating grouping structures into the prior improves the performance of R^2 decomposition priors, and under what conditions. To do so, we compare a baseline model M with its grouped counterpart M_G. For a given quantity of interest $\mathcal{Q}$, we define the performance difference as $\Delta\mathcal{Q} = \mathcal{Q}(M_G) - \mathcal{Q}(M)$. When

the distribution of $\Delta\mathcal{Q}$ is centered around zero, both models perform similarly. The interpretation of $\Delta\mathcal{Q}$ depends on the specific quantity: higher values may indicate improvement or deterioration.

We evaluate four aspects of model performance: predictive performance, parameter recovery, coverage, and convergence. Results on predictive performance and parameter recovery are reported in the main text while coverage and convergence diagnostics appear in the Appendix.

Out-of-sample Predictive Performance: We assess predictive performance using the expected log-pointwise predictive density (ELPD) (Vehtari and Ojanen, 2012; Vehtari et al., 2016), computed as

$$\text{ELPD} = \sum_{i=1}^{N_{\text{new}}} \ln \left(\frac{1}{S} \sum_{s=1}^{S} p(y_{\text{new},i} \mid \theta^{(s)}) \right), \tag{31}$$

where $\theta^{(s)}$ represents the sth draw from the posterior distribution $p(\theta \mid y)$ for $s = 1, \ldots, S$. The ELPD quantifies predictive performance across a set of N_{new} observations unseen during model training. A value of $\Delta\text{ELPD} > 0$ indicates that the grouped version outperforms the nongrouped. To account for variation in the scale of ΔELPD across different models and simulation settings, we report asinh-transformed ΔELPD values in all figures. This transformation is given by $\text{asinh}(x) = \log(x + \sqrt{x^2 + 1})$ and behaves like a logarithm for large values while remaining linear around zero, facilitating interpretation of differences.

Parameter Recovery: We assess accuracy of parameter recovery by the posterior root mean squared error (RMSE) for each regression coefficient:

$$\text{RMSE} = \frac{1}{p} \sum_{k=1}^{p} \sqrt{\frac{1}{S} \sum_{s=1}^{S} \left(b_k^{(s)} - b_k \right)^2}, \tag{32}$$

where $b_k^{(s)}$ is the sth posterior draw for coefficient b_k, and b_k is the true value. This global metric captures both systematic bias and estimation variance (Robert, 2007). To refine this assessment, we also report three types of RMSE: 1) across all coefficients, 2) for true zeros (evaluating shrinkage of irrelevant predictors), and 3) for true nonzeros (measuring signal recovery). Lower RMSE values indicate better recovery; thus, $\Delta\text{RMSE} < 0$ favors the grouped model. We highlight that this is not the typical RMSE based on the posterior mean, but we are using the whole posterior distribution to calculate it.

Coverage: We assess 95% marginal credible interval coverage by reporting average coverage rates, interval widths, specificity, sensitivity (power), and nonzero coverage (Benjamini and Hochberg, 1995). We also examine how these metrics vary with interval width via Receiver Operating Characteristic (ROC) curves.

Convergence Diagnostics: We monitor the quality of posterior samples using two convergence indicators: 1) The potential scale reduction factor $\widehat{R}$ (Vehtari et al., 2021), which should be close to 1 if chains are well mixed. 2) The effective sample size (ESS), reflecting the effect of dependency compared to a sample of independent draws. High ESS values signal efficient sampling and low autocorrelation (Brooks et al., 2011).

3.2 Simulations

Data generating procedure

We base our simulation design on a modified version of the setup proposed by Boss et al. (2023). The design matrix X is drawn from a multivariate normal distribution with zero mean and covariance matrix Σ_X. The structure of Σ_X reflects a block exchangeable correlation pattern: variables within the same group have pairwise correlation $\rho_{\text{in}} = 0.8$, while variables across groups have $\rho_{\text{out}} = 0.2$. All covariates are standardized to have unit variance. The error variance σ^2 is set to achieve a prespecified $R^2 \in \{0.25, 0.80\}$. We fix the sample size to $n = 200$ for all experiments. Each group contains $p_g = 10$ covariates, so the total number of groups is $G = p/10$, with $p \in \{100, 500\}$.

We evaluate five scenarios for generating the regression coefficients b, each reflecting different assumptions about sparsity and signal distribution within groups. 1) **Concentrated**: The signal is sparse and localized within each group. The first coefficient of each group is nonzero and set to $b_{g1} = 2$ while all others are zero. 2) **Random concentrated**: A random variant of the previous one in which the first coefficient in each group is sampled independently from $\mathcal{N}(0, 3^2)$, with all other coefficients in the group set to zero. 3) **Distributed**: The signal is spread across the multiple covariates that belong to a group. In the deterministic version, only the first group is active, leading to the first 10 coefficients being nonzero. We set $b_i = 0.5$ for $i = 1, \ldots, 5$ and $b_i = 1$ for $i = 6, \ldots, 10$. 4) **Random distributed**: Similar to the distributed, but the first 10 coefficients of the active group are independently drawn from $\mathcal{N}(0, 3^2)$. In both cases, the remaining coefficients are set to zero. 5) **Random Coefficients**: Following Boss et al. (2023), the first group is randomly assigned either a concentrated or distributed signal with equal probability, ensuring the presence of at least one active group. For each of the remaining groups, a concentrated signal is assigned with probability 0.2, a distributed signal with probability 0.2, and no signal with probability 0.6.

Models

The grouped and nongrouped versions differ only in their hierarchical structure: the grouped models include the ϕ and φ_g decompositions, while the nongrouped ones do not. For Group-R2 priors, we specify symmetric Dirichlet distributions for both ϕ and φ_g, with $a_G \in \{0.1, 0.5, 1\}$ and $c_g = 0.5$, respectively. For the nongrouped (standard) R2D2 model, we also use a symmetric Dirichlet with $a_\pi = a_G$.

We adopt the marginal formulations developed in Section 2.4 in Propositions 6 and 7 to set the hyperparameters of the R^2 prior. Specifically, we let $a_1 = Ga_G$, where G is the number of groups, and fix $a_2 = 0.5$ to induce heavy tails and facilitate signal detection. This approach enables an automatic and principled specification of the R^2 prior in line with the theoretical results. The value $c_g = 0.5$ was chosen based on the expected effective number of nonzero coefficients m_{eff}. As discussed in Section 2.5, it allows for a variety of shapes for its distribution, enforcing sparsity when a_G is small and a uniform like behavior that spreads signal across group elements when a_G increases. Additionally, Figure 4 shows that the chosen values for a_G when $c_g = 0.5$, allow for a

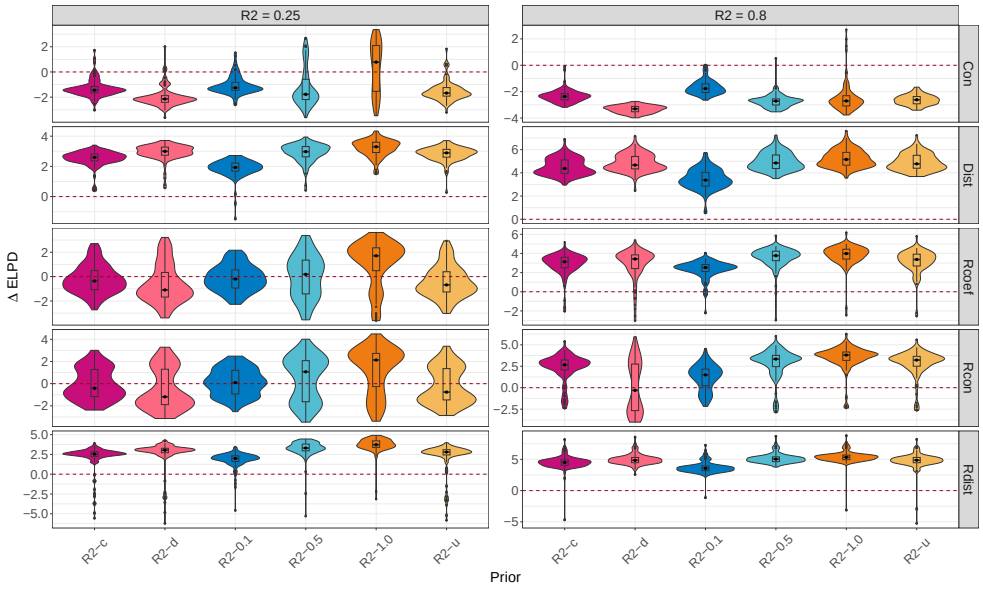


Figure 6: **ΔELPD (asinh-transformed) for the lower-dimensional scenario** ($p = 100$**,** $n = 200$**)**. Values above the line indicate improvement in predictive performance when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Group-R2 priors improve predictive performance in distributed and mixed signal settings. Gains increase with a_G, especially when R^2 is high.

wide range of correlation in log-marginal variances. A priori, these behaviors should accommodate for the different mechanisms by which we generate the coefficients. Since these models are fully characterized by the choice of a_G, we denote them as `R2-a_G` in the results.

We also evaluate two additional prior configurations and their grouped counterparts: 1) Uniform R2 prior `R2-u`: This baseline sets $a_1 = a_2 = 1$ for the R^2 prior, and uses symmetric Dirichlet distributions with $a_G = c_g = 1$ for the grouped version, or $a_\pi = 1$ for the nongrouped. This corresponds to uniform priors over both the total R^2 and the simplex partitions. 2) Concentrated and Distributed R2 Priors: These models share the same prior mean and variance for R^2, with $(\mu_{R^2}, \nu_{R^2}) = (1/3, 3)$. They differ in how signal sparsity is structured. The concentrated version `R2-c` sets $a_G = 1$ and $c_g = 0.5$, promoting sparsity within each group. This setup should be better suited for concentrated signals, where only a few predictors per group are truly active. In contrast, the distributed version `R2-d` uses $a_G = 0.5$ and $c_g = 1$, placing more uniform prior mass across coefficients within a group and thereby supporting distributed signals, where multiple predictors share the signal. These configurations allow us to test whether tuning

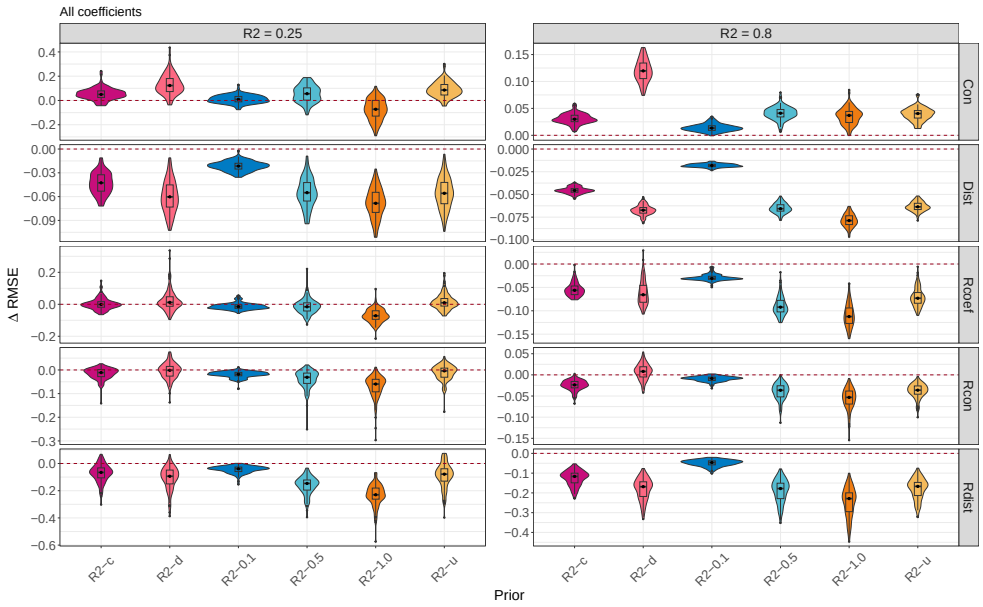


Figure 7: **ΔRMSE for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Group-R2 priors consistently improve parameter recovery in distributed and mixed signal settings. Improvements grow with a_G; the uniform prior performs robustly across different scenarios.

the decomposition hyperparameters to match the underlying sparsity pattern improves model performance.

Results

Lower-dimensional results We first evaluate the performance of Group-R2 priors in the lower-dimensional setting ($p = 100$, $n = 200$). Figure 6 reports the difference in out-of-sample predictive performance ΔELPD, while Figure 7 reports the difference in parameter recovery ΔRMSE. Overall, grouping improves predictive performance in scenarios where signal is distributed across covariates or randomly assigned (Distributed, Random Distributed, Random Coefficients), particularly when $R^2 = 0.8$. Gains in predictive performance tend to increase with the group-level concentration parameter a_G, as seen in the progression from models `R2-0.1` to `R2-1.0`. In contrast, when signals are highly concentrated within groups (Concentrated), grouping provides little to no benefit in predictive performance. Moreover, priors assuming distributed signals such

as `R2-d` perform worse in the concentrated scenario than priors assuming concentrated scenario like `R2-c`.

Figure 7 shows that, except for the concentrated "Con" case, parameter recovery is consistently improved when grouping is taken into account, especially when R^2 is high. Once again, concentrated signals represent the most challenging scenario, where non-grouped models can sometimes match or outperform grouped ones. As with predictive performance, RMSE improves monotonically with increasing a_G in the Group-R2 priors. The uniform prior `R2-u`, which does not favor any particular sparsity structure, emerges as a robust default. It delivers competitive performance across different signal patterns when no prior knowledge about sparsity is available. The improvement in the cases in which the grouped priors outperform, as can be seen from Figures 8 and 9, results from improved shrinkage on the truly zero coefficients groups. This holds true for low and high R^2. Relatively weak performance of the grouped priors in the concentrated case comes from the fact that the truly non-zero coefficients in each group are over-shrunk, as indicated by higher ΔRMSE seen in the first row of Figure 9. Taken together, these results echo findings found in the group-prior literature such as Boss et al. (2023), in which the benefit to using group-priors comes from imposing correlation in shrinkage for groups of covariates for which the signal is weak.

High-dimensional results We now consider the high-dimensional setting with $p = 500$ and $n = 200$. Figures 10 and 14 show ΔELPD and ΔRMSE, respectively[3].

The interpretation of the results remains the same. The results show similar patterns to the lower-dimensional case, but the benefits of incorporating group structure are amplified in high dimensions. In distributed signal settings, grouping leads to large gains in both predictive performance and parameter recovery, particularly when $R^2 = 0.8$. For the Random Coefficients case, improvement only comes when considering sufficiently high values for $a_G = 1$ and under high R^2. Concentrated signals are still challenging. Inclusion of grouping structures does not outperform the baseline.

3.3 Comparison with the GIGG prior

Among the grouped shrinkage priors discussed in Section 2.2, the GIGG prior of Boss et al. (2023) is the closest competitor to ours. It targets the same setting of block correlated regressors, it belongs to the same class of continuous shrinkage priors, and its three scales play the role that the two stage decomposition plays in the Group-R2 prior. Additionally it compares favorably against competing continuous as well as discrete mixture priors, extensively discussed in Boss et al. (2023).

The GIGG is also the only competitor for which a like for like comparison is possible. In particular, the GIGG prior employs a three tiered shrinkage prior $b_{gl} \sim \text{Normal}(0, \tau^2\gamma_g^2\lambda_{gl}^2)$ where $\tau^2 \sim \text{C}_+(0, 1)$, $\gamma_g^2 \sim \text{Gamma}(a_g, 1)$ and $\lambda_{gl}^2 \sim \text{InvGamma}(b_g, 1)$

[3]The break down of the ΔRMSE in terms of true zero and true non-zero is relegated to the Appendix due to similarity to the lower-dimensional DGP results.

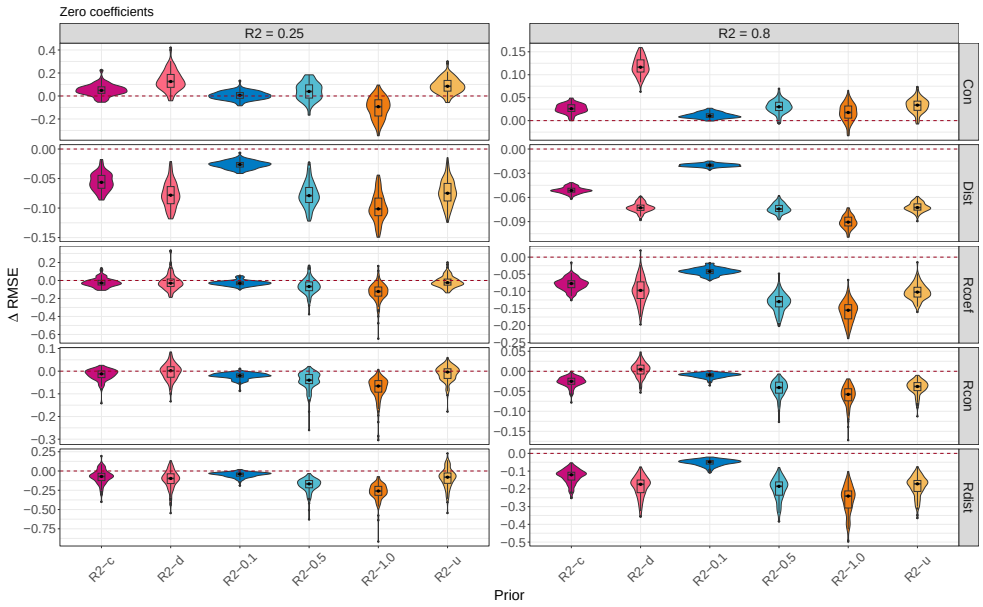


Figure 8: **ΔRMSE for truly zero coefficients for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

such that $\gamma_g^2\lambda_{gl}^2 \sim \text{BetaPrime}(a_g, b_g)$ and thus compares under the coupled hyperparameters suggested above to the Group-R2 in that $\tau_g^2 = \phi_g\tau^2 \sim \text{BetaPrime}(a_G, a_2)$, albeit the Group-R2 adds further decomposition by ϕ_{gl}. The two shapes in the GIGG prior (a_g, b_g) have distinct roles. The group shape a_g governs the mass that the marginal prior of b_{gl} places at the origin and therefore how strongly a group is thresholded as a whole, with smaller values shrinking entire groups more aggressively. The local shape b_g governs the tail of that same marginal and the degree to which coefficients within a group share their shrinkage, with larger values tying them more closely to the group scale and smaller values allowing individual coefficients to escape it. Since a major conceptual difference is direct control of the Group-R2 over the model's R^2, we suggest below a calibration method to find a_g and b_g to match the implied R^2 to the one set by our prior.

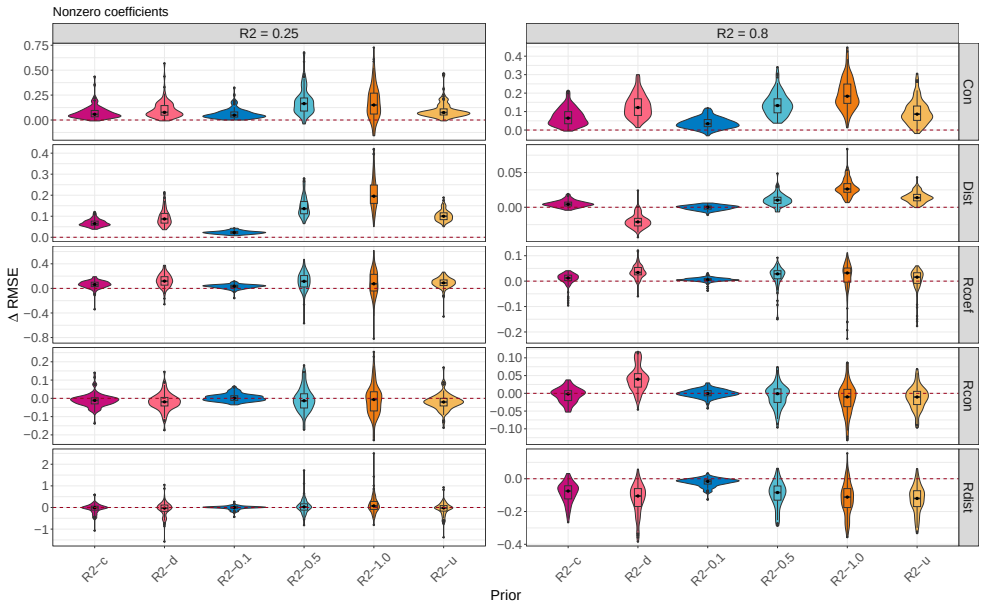


Figure 9: **ΔRMSE for truly nonzero coefficients for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

Calibrating the GIGG to the Group-R2 prior

Under the hierarchy of Section 2.2, and taking $\sigma^2 = 1$ and $X'X/n = I$ for the purpose of the calibration, the total prior signal variance under the GIGG is

$$V = \tau^2 \sum_{g=1}^{G} \gamma_g^2 \sum_{l=1}^{p_g} \lambda_{gl}^2, \qquad R^2_{\text{GIGG}} = \frac{V}{V+1}, \tag{33}$$

whose induced distribution depends on (a_g, b_g) but has no closed form. The Group-R2 prior with concentration a_G places $R^2 \sim \text{Beta}(Ga_G, a_2)$ directly, with $a_2 = 0.5$ as above. For each $a_G \in \{0.1, 0.5, 1\}$ we therefore choose

$$(a_g, b_g) = \underset{(a_g, b_g)}{\arg\min}\; W_1\big(R^2_{\text{GIGG}}(a_g, b_g),\; \text{Beta}(Ga_G, a_2)\big)\,, \tag{34}$$

where W_1 is the 1-Wasserstein distance between the two distributions of R^2. We denote the resulting models `GIGG-`a_G and pair each with the Group-R2 model `R2-`$\mathtt{a_G}$ that shares its target. Both sides of Equation (34) move with the number of groups, so the calibration

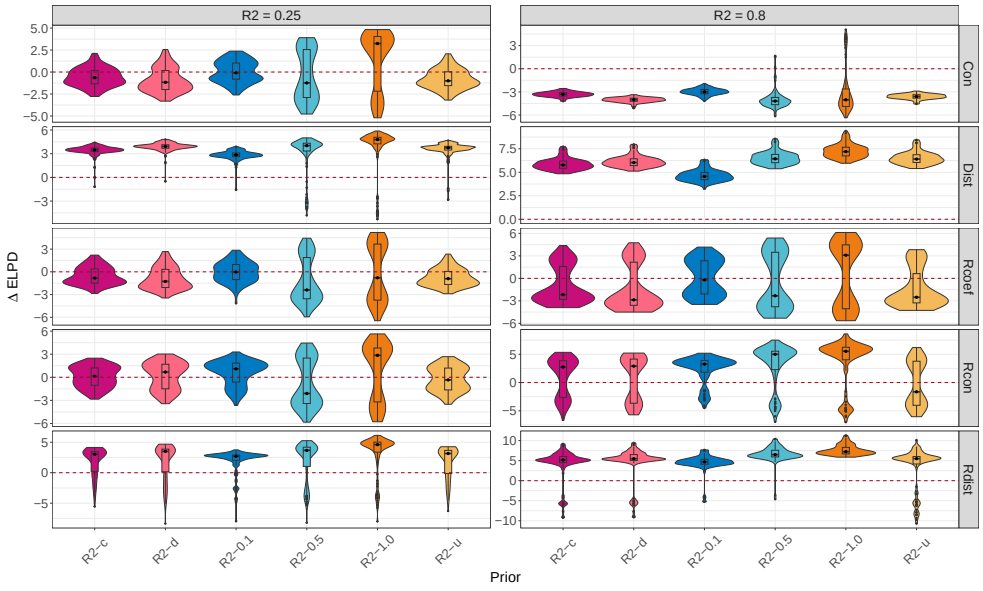


Figure 10: **ΔELPD (asinh-transformed) for the high-dimensional scenario ($p = 500$, $n = 200$).** Values above the line indicate improvement in predictive performance when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Group-R2 priors improve predictive performance in distributed signal settings. Performance in Random Coefficients (Rcoef) is mixed, with limited gains except when a_G is sufficiently high and high R^2 is present. Concentrated signals remain challenging.

is carried out separately for $G = 10$ and $G = 50$. The optimisation, its constraints and its diagnostics are described in Section 5.5 of the Appendix.

Table 1 reports the calibrated values. The reported W_1 throughout, but two qualifications apply. At $G = 10$ the constraint $a_g \leq 0.5$, which keeps the pole of the marginal prior at the origin (Boss et al., 2023), binds for `GIGG-0.5` and `GIGG-1.0`, so these two are matched only as closely as the sparse regime allows. At $G = 50$ no constraint binds, but the targets are themselves close to degenerate, since $\text{Beta}(50, 0.5)$ places almost all of its mass above 0.98, and the small W_1 reflects both distributions sitting near the boundary rather than agreement in shape. Figure 26 in the Appendix shows the two distributions of R^2 side by side for each design and makes both qualifications visible.

Table 1: GIGG shapes calibrated to the R^2 prior implied by each Group-R2 model, with the out of sample 1-Wasserstein distance to the target. Values marked † sit at the boundary of the feasible region.

	$G = 10$ ($p = 100$)			$G = 50$ ($p = 500$)		
Model	a_g	b_g	W_1	a_g	b_g	W_1
`GIGG`-0.1	0.363	$10.000^\dagger$	0.038	0.488	3.826	0.058
`GIGG`-0.5	$0.500^\dagger$	1.394	0.058	0.051	0.393	0.020
`GIGG`-1.0	$0.500^\dagger$	0.714	0.042	0.029	0.201	0.010

Results

Figure 11 reports $\Delta\text{ELPD} = \text{ELPD}(\texttt{R2-a}_\texttt{G}) - \text{ELPD}(\texttt{GIGG-}a_G)$ for the lower-dimensional setting, so that positive values favor the Group-R2 prior. The corresponding high-dimensional figure and the ΔRMSE figures, both over all coefficients and split by coefficient type, are relegated to Section 5.5 of the Appendix.

The comparison depends strongly on a_G. At $a_G = 0.1$ the Group-R2 prior predicts better, with a median ΔELPD of 5.4 and an advantage in 69% of replications. At $a_G = 0.5$ and $a_G = 1$ the ordering reverses, with medians of -7.8 and -28.6. The same ordering holds in the high-dimensional setting ($p = 500$, Figure 27 in the Appendix), where the medians are -7.6, -184.7 and -357.7. Parameter recovery follows the same pattern, with median ΔRMSE of -0.069, 0.027 and 0.180 at $p = 100$, shown together with the high-dimensional case in Figures 28 and 29 of the Appendix.

One possible contributor to this ordering is the rule $a_1 = Ga_G$ used to set the R^2 prior. Since $a_2 = 0.5$ is fixed, the prior mean of R^2 is $Ga_G/(Ga_G + 0.5)$, which is 0.67, 0.91 and 0.95 at $G = 10$ and 0.91, 0.98 and 0.99 at $G = 50$. The data are generated with $R^2 \in \{0.25, 0.80\}$, so for the larger values of a_G the prior places almost all of its mass above the truth, and increasingly so as G grows. For $a_G = 0.1$ the median ΔELPD is 14.6 when $R^2 = 0.80$ and -0.6 when $R^2 = 0.25$ at $p = 100$, and 39.8 against -20.4 at $p = 500$, so the Group-R2 prior is favored precisely where its implied prior mean is close to the data generating value. The calibrated GIGG inherits the same target, but the mapping from (a_g, b_g) to R^2 in Equation (33) is indirect enough that it does not reproduce the same concentration, and the resulting prior is less informative about the total explained variance and more robust to this misspecification.

The difference in recovery is concentrated in the coefficients that are truly zero. Splitting ΔRMSE by coefficient type, the median difference on the truly nonzero coefficients stays close to zero in every configuration, whereas on the truly zero coefficients it grows with a_G and with the dimension and accounts for most of the aggregate difference. The split figures are given in Section 5.5 of the Appendix.

The quality of the posterior sampling qualifies these numbers, and convergence diagnostics for all six models, split by coefficient type and by scenario, are reported in Tables 6 and 7 of the Appendix. The posterior geometry induced by the GIGG is the more difficult of the two, and increasingly so as a_G grows and as the dimension increases.

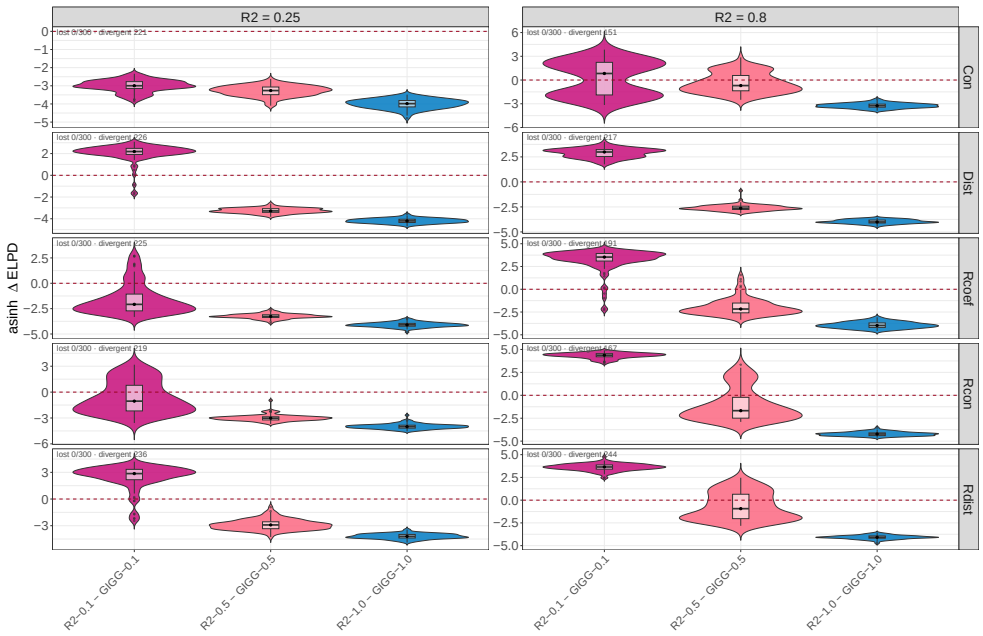


Figure 11: **ΔELPD (asinh-transformed) against the calibrated GIGG for the lower-dimensional scenario** ($p = 100$**,** $n = 200$). Values above the line indicate better predictive performance of the Group-R2 prior. Each contrast pairs a Group-R2 model with the GIGG calibrated to the R^2 prior it implies, so that the two differ only in how the explained variance is decomposed. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. The panel annotations give the number of contrasts lost to failed fits and the number of retained contrasts in which at least one of the two fits had divergent transitions.

In the high-dimensional setting the Group-R2 models attain higher effective sample sizes and lower $\widehat{R}$ than their calibrated counterparts at $a_G \in \{0.5, 1\}$, and for `GIGG-1.0` the mixing is poor enough that the posterior cannot be regarded as adequately explored.

Taken together, the Group-R2 compares favorably to the GIGG when the R2 prior is well calibrated to the true R^2 and offers the clear advantage of interpretability of the global scale being matched to a prior on the models' R^2 and control over the decomposition.

4 Discussion

In this work, we introduced the Group-R2 prior, analyzed its mathematical properties, and evaluated its empirical performance through simulations. The Group-R2 prior is an extension of the R2-based shrinkage priors that incorporates known grouping structures through a two-stage hierarchical decomposition of the prior coefficient of determination R^2. We first distribute the total variance across groups and then within each group across individual coefficients. We have shown that this construction preserves desirable shrinkage properties, such as heavy tails and strong concentration near the origin, both at a marginal and group level. We have also shown how the choice of hyperparameters of the decomposition governs shrinkage behavior at both levels of the hierarchy. We have also provided a practical framework for hyperparameter calibration using interpretable diagnostics such as the effective number of non-zero coefficients, with recommendations on how to further tune hyperparameters based on knowledge of R^2 or the type of signal expected from a given group.

Our experiments complement our theoretical results and illustrate the practical value of incorporating grouping structures into R2-based priors. In both lower- and high-dimensional settings, Group-R2 priors consistently improve parameter recovery when the signal is distributed across predictors or randomly mixed. In line with our discussion, increasing a_G enhances recovery by encouraging more even allocations of variance across groups. Gains in predictive accuracy follow similar trends, especially in high dimensions. When the signal was concentrated in a single coefficient within each group, grouping brought little or no benefit and could make recovery worse, depending on how well the prior matched the data generating process. However, we conjecture that different hyperparameter specification might lead to improvements in this scenario.

Beyond the comparison with non-grouped R2-based priors, we compared the Group-R2 prior with a GIGG prior calibrated to the same implied distribution of R^2, and with several grouped priors from the literature, including spike-and-slab formulations, the group horseshoe, GRASP and a grouped g-prior. The Group-R2 prior was competitive in these comparisons across the different settings considered, particularly when the prior on R^2 was close to the explained variance in the data. When the prior placed most of its mass near the true value, the Group-R2 prior showed excellent performance. This showcases that R2 elicitation is a crucial step in R2-based priors. In applications we therefore recommend choosing the location of the R^2 prior from domain knowledge.

A key limitation of the current framework is its reliance on pre-specified group structures, which are assumed to be known and fixed in advance. In practice, groups may be misspecified due to noisy prior knowledge or arbitrary definitions, leading to misallocation of variance and suboptimal shrinkage. For example, assigning signal covariates to low-variance groups can suppress important effects, while allocating variance to purely noisy groups may inflate uncertainty. This makes the method vulnerable to structural assumptions, particularly in settings where group definitions are uncertain or high-dimensional. Future work should focus on extending the Group-R2 framework to settings where group structure is unknown or partially observed, for instance by coupling with latent group discovery methods or using external similarity measures to inform grouping.

Acknowledgments

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project 500663361. Javier Enrique Aguilar and Paul-Christian Bürkner acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by DFG Project 271512359. Paul-Christian Bürkner further acknowledges support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the Collaborative Research Center 391 (Spatio-Temporal Statistics for the Transition of Energy and Transport) – 520388526. Javier Enrique Aguilar further acknowledges travel support from the European Union's Horizon 2020 research and innovation programme under grant agreement No 951847. David Kohns acknowledges the computational resources provided by the Aalto Science-IT project, and the support of the Research Council of Finland Flagship programme: Finnish Center for Artificial Intelligence, Research Council of Finland project (340721), and the Finnish Foundation for Technology Promotion.

5 Appendix

5.1 Proofs

Proof of Proposition 1

Recall that $R_g^2 = \phi_g R^2$, where ϕ and R^2 are independent. Since the beta-distributed variable R^2 is positive almost surely, division by R^2 gives

$$\left(\frac{R_1^2}{R^2}, \ldots, \frac{R_G^2}{R^2}\right) = (\phi_1, \ldots, \phi_G) \sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_G). \tag{35}$$

Write $\alpha_0 = \sum_g \alpha_g$. To obtain the distribution of a single coordinate, represent the Dirichlet vector as $\phi_g = Y_g / \sum_h Y_h$, using independent gamma variables $Y_h \sim \text{Gamma}(\alpha_h, 1)$, with shape and rate parameterization. Then $\sum_{h \neq g} Y_h$ is gamma with shape $\alpha_0 - \alpha_g$, independently of Y_g. The ratio of these two gamma components to their sum has the beta distribution, so

$$\frac{R_g^2}{R^2} = \phi_g \sim \text{Beta}(\alpha_g, \alpha_0 - \alpha_g). \tag{36}$$

Independence ensures that conditioning on $R^2 = r^2$ leaves the law of ϕ unchanged. The conditional group contributions are therefore a Dirichlet vector multiplied by r^2. For distinct g, h, the Dirichlet moments give (Lin, 2016)

$$\begin{aligned} \mathbb{E}(\phi_g) = \frac{\alpha_g}{\alpha_0}, \qquad \mathbb{E}(\phi_g \phi_h) &= \frac{\alpha_g \alpha_h}{\alpha_0(\alpha_0 + 1)}, \\ \text{Cov}(R_g^2, R_h^2 \mid R^2 = r^2) = (r^2)^2 \left\{ \frac{\alpha_g \alpha_h}{\alpha_0(\alpha_0 + 1)} - \frac{\alpha_g \alpha_h}{\alpha_0^2} \right\} &= -\frac{(r^2)^2 \alpha_g \alpha_h}{\alpha_0^2(\alpha_0 + 1)}. \end{aligned} \tag{37}$$

This proves the three assertions of the proposition. □

Proof of Proposition 2

To prove the proposition we will use the following auxiliary propositions. Their proofs can be found in Zhang et al. (2020).

Proposition 12. *If* $\tau^2 \mid \xi \sim Ga(a_1, \xi)$ *and* $\xi \sim Ga(a_2, 1)$*, then* $\tau^2 \sim \text{BetaPrime}(a_1, a_2)$.

Proposition 13. *If* $\tau^2 \mid \xi \sim Ga(a_1, \xi)$, $\phi \sim \text{Dirichlet}\, s(a_\pi)$ *of dimension* k *and* $a_1 = k a_\pi$*, then* $\phi_j \tau^2 \mid \xi \sim Ga(a_\pi, \xi), j = 1, ..., k$ *independently.*

1. By hypothesis we have that $\tau^2 \sim \text{BetaPrime}(Ga_G, a_2)$. Applying Proposition 12 yields

   $$\tau^2 \mid \xi \sim Ga(Ga_G, \xi), \xi \sim Ga(a_2, 1).$$

   Using Proposition 13 with $a_1 = Ga_G$ shows that $\tau_g^2 \mid \xi \sim Ga(a_G, \xi)$. Therefore we have

$$\tau_g^2 \mid \xi \sim Ga(a_G, \xi), \xi \sim Ga(a_2, 1).$$

Applying Proposition 12 again gives $\tau_g^2 \sim \text{BetaPrime}(a_G, a_2)$. □

2. The result follows by applying the same procedure that we have used to show the previous result but at a group level. □

Proof of Proposition 3

We construct both allocation levels from one family of independent gamma variables. Let $Z_{gl} \sim \text{Gamma}(c_{gl}, 1)$ independently, and define

$$H_g = \sum_{l=1}^{p_g} Z_{gl}, \qquad H_+ = \sum_{g=1}^{G} H_g, \qquad a_g^* = \sum_{l=1}^{p_g} c_{gl}. \tag{38}$$

Each H_g is gamma with shape a_g^*. The different group sums are independent. Within group g, gamma normalization gives

$$\left(\frac{Z_{g1}}{H_g}, \ldots, \frac{Z_{gp_g}}{H_g}\right) \sim \text{Dirichlet}(c_{g1}, \ldots, c_{gp_g}), \tag{39}$$

independently of H_g. Since different groups use disjoint independent gamma variables, the within-group vectors are mutually independent and independent of the entire vector of group sums. Therefore we may set

$$\phi_g = H_g/H_+, \qquad \varphi_{gl} = Z_{gl}/H_g. \tag{40}$$

This construction has the group Dirichlet distribution and all the independence properties in the proposition. The product of the two allocations simplifies to

$$\phi_g \varphi_{gl} = \frac{Z_{gl}}{H_+} = \frac{Z_{gl}}{\sum_{h,r} Z_{hr}}. \tag{41}$$

The concatenated product vector is the normalization of independent gamma variables with shapes $(c_{gl})_{g,l}$. Its distribution is the stated Dirichlet distribution. In the case of singleton group, their within-group proportion equals one. □

Proof of Proposition 4

1. We are interested in the quantity

$$\text{Corr}\left(\log(\phi_g\varphi_{gj}), \log(\phi_g\varphi_{gk})\right) = \frac{\text{Cov}\left(\log(\phi_g\varphi_{gj}), \log(\phi_g\varphi_{gk})\right)}{\sqrt{\text{Var}(\log(\phi_g\varphi_{gj})\,\text{Var}(\log(\phi_g\varphi_{gk})}}, \quad j \neq k \tag{42}$$

Applying properties of the covariance operator and using the fact that ϕ_g and φ_g are independent we find that:

$$\begin{aligned}\text{Cov}(\log(\phi_g\varphi_{gj}), \log(\phi_g\varphi_{gk})) &= \text{Cov}\left(\log(\phi_g) + \log(\varphi_{gj}), \log(\phi_g) + \log(\varphi_{gk})\right) \\ &= \text{Var}(\log\phi_g) + \text{Cov}(\log\varphi_{gj}, \log\varphi_{gk}).\end{aligned} \tag{43}$$

The main quantity in the denominator is $\text{Var}(\log(\phi_g\varphi_{gj})$, which can be written as

$$\begin{aligned}\text{Var}(\log(\phi_g\varphi_{gj}) &= \text{Var}\left(\log(\phi_g) + \log(\varphi_{gj})\right) \\ &= \text{Var}(\log(\phi_g)) + \text{Var}(\log(\varphi_{gj}).\end{aligned} \tag{44}$$

Since $\varphi_g \sim \text{Dirichlet}\, s(c_g)$, both φ_{gj} and φ_{gk} have the same marginal distributions and we have $\text{Var}(\log(\phi_g\varphi_{gj}) = \text{Var}(\log(\phi_g) + \text{Var}(\log(\varphi_{g\cdot})$. Therefore

$$\text{Corr}\left(\log(\phi_g\varphi_{gj}), \log(\phi_g\varphi_{gk})\right) = \frac{\text{Var}(\log\phi_g) + \text{Cov}(\log\varphi_{gj}, \log\varphi_{gk})}{\text{Var}(\log(\phi_g) + \text{Var}(\log(\varphi_{g\cdot})}. \quad \square \tag{45}$$

The result can be generalized to the case in which the proportions of explained variance do not follow a symmetric Dirichlet, but we do not show that here.

2. The result follows by substituting the following quantities in Equation (45):

$$\begin{aligned}\text{Var}(\log\phi_g) &= \psi_1(a_g) - \psi_1(Ga_g) \\ \text{Var}(\log\varphi_{gl}) &= \psi_1(c_g) - \psi_1(p_gc_g), \quad l \in \{1, ..., p_g\} \\ \text{Cov}(\log\varphi_{gj}, \log\varphi_{gk}) &= -\psi_1(p_gc_g), \quad j \neq k\end{aligned} \tag{46}$$

where ψ_1 is the trigamma function (Olver et al., 2010b). $\square$

3. We have

$$\begin{aligned}\text{Cov}(\phi_g\varphi_{gj},\ \phi_g\varphi_{gk}) &= \mathbb{E}[\phi_g^2\varphi_{gj}\varphi_{gk}] - \mathbb{E}[\phi_g\varphi_{gj}]\mathbb{E}[\phi_g\varphi_{gk}] \\ &= \mathbb{E}[\phi_g^2]\cdot\mathbb{E}[\varphi_{gj}\varphi_{gk}] - \mathbb{E}[\phi_g]^2\mathbb{E}[\varphi_{gj}]\mathbb{E}[\varphi_{gk}], \quad j \neq k,\end{aligned} \tag{47}$$

where we have used the independence of ϕ_g, φ_g to expand the expectations. The expressions in the last line have closed analytical forms given by:

$$\begin{aligned}&\mathbb{E}[\phi_g] = \frac{1}{G}, \quad \mathbb{E}[\varphi_{gl}] = \frac{1}{p_g}, \quad l \in \{1, ..., p_g\}, \\ &\mathbb{E}[\phi_g^2] = \frac{a_g(a_g+1)}{(Ga_g)(Ga_g+1)} \quad \mathbb{E}[\varphi_{gj}\varphi_{gk}] = \frac{c_g}{p_g(c_gp_g+1)}, \quad j \neq k\end{aligned} \tag{48}$$

The result follows by substituting the expressions in Equation (47). $\square$

Proof of Proposition 5

Under the assumptions considered, the Group-R2 prior has the following alternative representation

$$b_{gl} \mid \lambda_{gl}^2 \sim \text{Normal}\left(0, \lambda_{gl}^2\right) \quad \lambda_{gl}^2 \sim \text{BetaPrime}(c_g, a_2). \tag{49}$$

The prior marginal density of b_i is given by

$$\begin{aligned} p(b_{gl}) &= \int_0^\infty \frac{1}{\sqrt{2\pi\lambda_{gl}}^2} \exp\left\{-\frac{b_{gl}^2}{2\lambda_{gl}^2}\right\} \frac{1}{\mathrm{B}(c_g, a_2)} (\lambda_{gl}^2)^{c_g-1} \left(\lambda_{gl}^2+1\right)^{-c_g-a_2} d\lambda_{gl}^2 \\ &= \frac{1}{\sqrt{2\pi}\,\mathrm{B}(c_g, a_2)} \int_0^\infty \exp\left\{-\frac{b_{gl}^2}{2} t_{gl}\right\} t_{gl}^{\eta-1} (t_{gl}+1)^{\nu-\eta-1} dt_{gl} \\ &= \frac{1}{\sqrt{2\pi}\,\mathrm{B}(c_g, a_2)} \Gamma(\eta) U(\eta, \nu, z_{gl}), \end{aligned} \tag{50}$$

where $t_{gl} = \frac{1}{\lambda_i^2}$, $\eta = a_2 + 1/2$ and $\nu = 3/2 - c_g$. $U(\eta, \nu, z_{gl})$ represents the confluent hypergeometric function of the second kind (Jeffrey et al., 2007)) and $z_{gl} = \frac{|b_{gl}|^2}{2}$. The function $U(\eta, \nu, z_{gl})$ is defined as long as the real part of η and z_{gl} are positive, which is always the case since $a_2 > 0$ and $|b_{gl}| > 0$. □

Proof of Proposition 6

In the following we drop the index of z_{gl}, since all cases are handled equally. The confluent hypergeometric function of the second kind $U(\eta, \nu, z)$ satisfies the following identity Olver et al. (2010a)).

$$U(\eta, \nu, z) = z^{1-\nu} U(\eta - \nu + 1, 2 - \nu, z). \tag{51}$$

Substituting $\eta = a_2 + 1/2, \nu = 3/2 - c_g$ and $z = \frac{t^2}{2}$ results in

$$U\left(a_2 + 1/2, 3/2 - c_g, \frac{t^2}{2}\right) = 2^{1/2-c_g} |t|^{2c_g-1} U\left(a_2 + c_g, c_g + 1/2, \frac{t^2}{2}\right). \tag{52}$$

The right hand side has a singularity at $t = 0$ when $c_g < 1/2$. We consider the limiting form U has as $|z| \to 0$ to study the order of concentration near the origin for different values of c_g. Equation 13.2.18 in Olver et al. (2010a) states

$$U(\eta, \nu, z) = \frac{\Gamma(\nu-1)}{\Gamma(\eta)} z^{1-\nu} + \frac{\Gamma(1-\nu)}{\Gamma(\eta-\nu+1)} + \mathcal{O}\left(z^{2-\nu}\right), \quad 1 < \nu < 2. \tag{53}$$

Substituting $\eta = a_2 + 1/2, \nu = 3/2 - c_g$ and $z = \frac{t^2}{2}$ results in

$$U\left(a_2 + 1/2, 3/2 - c_g, \frac{t^2}{2}\right) = C t^{2c_g-1} + \frac{\Gamma(c_g + 1/2)}{\Gamma(a_2 + 1/2)} + \mathcal{O}\left(|t|^{1+2c_g}\right), \quad 0 < c_g < 1/2, \tag{54}$$

where $C = 2^{1/2-c_g} \frac{\Gamma(1/2-c_g)}{\Gamma(a_2+c_g)}$. The last expression is unbounded at $t = 0$ when $c_g < 1/2$. The case when $c_g = 1/2$ follows equation 13.2.19 in Olver et al. (2010a) given by

$$U(\eta, \nu, z) = -\frac{1}{\Gamma(\eta)} \left(\ln(z) + \psi(\eta) + 2\gamma\right) + \mathcal{O}\left(z \ln(z)\right), \quad \nu = 1, \tag{55}$$

where $\psi(\cdot), \gamma$ represent the digamma function and the Euler-Mascheroni constant respectively. Substituting $\eta = a_2 + 1/2, \nu = 3/2 - c_g$ and $z = \frac{t^2}{2}$ gives

$$U\left(a_2 + 1/2, 3/2 - c_g, \frac{t^2}{2}\right) = -\frac{1}{\Gamma(a_2 + 1/2)}\left(\ln\left(\frac{t^2}{2}\right) + \psi(a_2 + 1/2) + 2\gamma\right) + \mathcal{O}\left(t^2 \ln\left(\frac{t^2}{2}\right)\right), \quad c_g = 1/2. \tag{56}$$

The last expression is not defined at $t = 0$ and goes to infinity as $|t| \to 0$. To see the marginal priors are bounded when $c_g > 1/2$ it is sufficient to consider the cases given by equations 13.2.20-13.2.22 in Olver et al. (2010a) and making the proper substitutions of $\eta = a_2 + 1/2, \nu = 3/2 - c_g$ and $z = \frac{t^2}{2}$. In terms of the values c_g takes we have

$$U(\eta, \nu, t) \sim \begin{cases} \mathcal{O}\left(t^{2c_g - 1}\right), & 1/2 < c_g < 3/2 \\ \mathcal{O}\left(t^2 \ln(t^2/2)\right), & c_g = 3/2 \\ \mathcal{O}\left(t^2\right), & c_g > 3/2 \end{cases} \tag{57}$$

The derivative of $U(\eta, \nu, z)$ is given by equation 13.3.22 in Olver et al. (2010a) as

$$\frac{d}{dz}U(\eta, \nu, z) = -\eta U(\eta + 1, \nu + 1, z). \tag{58}$$

Combining this with Equation (51) gives

$$\frac{d}{dz}U(\eta, \nu, z) = -\eta z^{-\nu} U(\eta - \nu + 1, 1 - \nu, z). \tag{59}$$

Substituting $\eta = a_2 + 1/2, \nu = 3/2 - c_g$ and $z = \frac{t^2}{2}$ results in

$$\frac{d}{dt}U\left(a_2 + 1/2, 3/2 - c_g, \frac{t^2}{2}\right) = Ct^{2c_g - 2}U\left(a_2 + c_g, c_g - 1/2, \frac{t^2}{2}\right), \tag{60}$$

where $C = -(a_2 + 1/2)2^{3/2 - c_g}$. Equation (60) is undefined at $t = 0$ when $c_g < 1$. □

Proof of Proposition 7

To prove the proposition we make use of Watson's lemma (Miller, 2006).

Proposition 14 (Watson's lemma)**.** *Let $0 \leq T \leq \infty$ be fixed. Assume $f(t) = t^\lambda g(t)$, where $g(t)$ has an infinite number of derivatives in the neighborhood of $t = 0$, with $g(0) \neq 0$, and $\lambda > -1$. Suppose, in addition, either that $|f(t)| < Ke^{ct}$ for any $t > 0$, where K and c are independent of t. Then it is true that for all positive x that*

$$\left|\int_0^T e^{-xt} f(t) dt\right| < \infty \tag{61}$$

and that the following asymptotic equivalence holds:

$$\int_0^T e^{-xt} f(t)dt \sim \sum_{n=0}^{\infty} \frac{g^{(n)}(0)\Gamma(\lambda+n+1)}{n!x^{\lambda+n+1}}, \tag{62}$$

for $x > 0$ *as* $x \to \infty$.

The marginal distribution of a coefficient b is given by

$$\begin{aligned} p(b) &= \frac{1}{\sqrt{2\pi}\,\mathrm{B}(c_g, a_2)} \int_0^{\infty} \exp\left\{-\frac{|b|^2}{2}t\right\} t^{\eta-1}(t+1)^{\nu-\eta-1}dt \\ &= \int_0^{\infty} \exp\{-zt\}\, f(t)dt, \end{aligned} \tag{63}$$

where $z = \frac{|b|^2}{2}, f(t) = C\,t^{\eta-1}(t+1)^{\nu-\eta-1} = t^{\eta-1}g(t), C = \left(\sqrt{2\pi}\mathrm{B}(c_g, a_2)\right)^{-1}$, and $g(t) = C\,(t+1)^{\nu-\eta-1}$. Setting $\lambda = \eta - 1$, the hypothesis $\lambda > -1$ is satisfied since $a_2 - 1/2 > -1$ for $a_2 > 0$. Moreover, $g(t)$ is infinitely differentiable around $t = 0$ and $g(0) \neq 0$. By Watson's Lemma, since $|f(t)| < Ke^{ct}$ for all $t > 0$ where K and c are independent of t, then as $|b| \to \infty$,

$$p(z \mid \sigma) = \sum_{n=0}^{\infty} \frac{g^{(n)}(0)\Gamma(\lambda+n+1)}{n!z^{\lambda+n+1}}. \tag{64}$$

Truncating the sum at $n = 2$ gives

$$\begin{aligned} p(z) = C\Bigg\{ &\frac{\Gamma(a_2+1/2)}{z^{a_2+1/2}} - \frac{(c_g+a_2)\Gamma(a_2+3/2)}{z^{a_2+3/2}} + \frac{(c_g+a_2)(c_g+a_2+1)\Gamma(a_2+5/2)}{z^{a_2+5/2}} \Bigg\} \\ &+ \mathcal{O}\left(\frac{1}{z^{a_2+7/2}}\right) \\ \sim\ &\mathcal{O}\left(\frac{1}{z^{a_2+1/2}}\right). \end{aligned} \tag{65}$$

Therefore $p\,(|b|) \sim \mathcal{O}\left(\frac{1}{|b|^{2a_2+1}}\right)$. When $a_2 < 1/2$ as $|b| \to \infty$ and comparing with a Cauchy distribution represented by $\frac{1}{b^2}$, we have that

$$\frac{p(b)}{\frac{1}{b^2}} \sim \mathcal{O}\left(\frac{1}{b^{2a_2-1}}\right) \to \infty. \tag{66}$$

Proof of Proposition 8

Since we have shown that $\lambda_{gl}^2 \sim \mathrm{BetaPrime}(c_g, a_2)$, we can write the Group-R2 prior as

$$b_{gl} \mid \lambda_{gl}^2 \sim \mathrm{Normal}(0, \lambda_{gl}^2), \quad \lambda_{gl}^2 \sim \mathrm{BetaPrime}\,(c_g, a_2)\,. \tag{67}$$

When $c_g = a_2 = 1/2$, then $\lambda_{gl} \sim \text{Cauchy}^+(0,1)$ and we have

$$b_{gl} \mid \lambda_{gl}^2 \sim \text{Normal}(0, \lambda_{gl}^2), \quad \lambda_{gl} \sim \text{Cauchy}^+(0,1). \tag{68}$$

The horseshoe prior has the following representation (Carvalho et al., 2010):

$$b_{gl} \mid \lambda_{gl}, \omega \sim \text{Normal}(0, \lambda_{gl}\omega), \quad \lambda_{gl} \sim \text{Cauchy}^+(0,1), \tag{69}$$

where λ_{gl} are per coefficient shrinkage parameters and ω is the global shrinkage parameter. Therefore the Group-R2 prior with $c_g = a_2 = \frac{1}{2}$ corresponds to a horseshoe prior with global scale $\omega = 1$. This proves the first part of the proposition.

To see that bounded influence holds, consider the scalar normal-means model $y \mid b \sim \text{Normal}(b, 1)$. Carvalho et al. (2010) shows the following posterior-mean identity

$$\mathbb{E}(b \mid y) - y = \frac{m'(y)}{m(y)}. \tag{70}$$

We call the right hand part of the equation the score. Since $y \mid \lambda \sim \text{Normal}(0, 1+\lambda^2)$, integrating over the half-Cauchy density gives

$$\begin{aligned} m(y) &= \frac{2}{\pi\sqrt{2\pi}} \int_0^\infty \exp\left\{-\frac{y^2}{2(1+\lambda^2)}\right\} (1+\lambda^2)^{-3/2}\, d\lambda. \\ &= \frac{1}{\pi\sqrt{2\pi}} \int_0^1 e^{-y^2 z/2} (1-z)^{-1/2}\, dz. \end{aligned} \tag{71}$$

where we have used the change of variables $z = (1+\lambda^2)^{-1}$. To evaluate this integral, set $u = \sqrt{1-z}$. Define the imaginary error function by

$$\text{erfi}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{t^2}\, dt. \tag{72}$$

Then, for $y \neq 0$,

$$\begin{aligned} m(y) &= \frac{2e^{-y^2/2}}{\pi\sqrt{2\pi}} \int_0^1 e^{y^2 u^2/2}\, du \\ &= \frac{e^{-y^2/2}}{\pi y} \text{erfi}\left(\frac{y}{\sqrt{2}}\right). \end{aligned} \tag{73}$$

At $y = 0$, the preceding integral gives $m(0) = 2/(\pi\sqrt{2\pi})$. Differentiating for $y \neq 0$, using $\text{erfi}'(x) = 2e^{x^2}/\sqrt{\pi}$, gives

$$\frac{m'(y)}{m(y)} = -y - \frac{1}{y} + \frac{\sqrt{2/\pi}\, e^{y^2/2}}{\text{erfi}(y/\sqrt{2})}. \tag{74}$$

We now evaluate the tail limit of this expression. The error-function asymptotic expansion (Olver et al., 2010b), applied to $\text{erfi}(x) = -i\,\text{erf}(ix)$, gives

$$\text{erfi}\left(\frac{y}{\sqrt{2}}\right) = \frac{\sqrt{2/\pi}\, e^{y^2/2}}{y} \left\{1 + \frac{1}{y^2} + O(y^{-4})\right\}, \qquad y \to +\infty. \tag{75}$$

Hence, the last term of Equation (74) satisfies

$$\frac{\sqrt{2/\pi}\, e^{y^2/2}}{\operatorname{erfi}(y/\sqrt{2})} = y\left\{1 + \frac{1}{y^2} + O(y^{-4})\right\}^{-1} = y - \frac{1}{y} + O(y^{-3}). \tag{76}$$

Substituting into Equation (74) gives

$$\frac{m'(y)}{m(y)} = -\frac{2}{y} + O(y^{-3}). \tag{77}$$

Since m is even, $\frac{m'(y)}{m(y)}$ is odd and the expansion holds in both tails. Therefore

$$\mathbb{E}(b \mid y) - y = \frac{m'(y)}{m(y)} = -\frac{2}{y} + O(|y|^{-3}) \longrightarrow 0, \qquad |y| \to \infty. \tag{78}$$

It remains to check boundedness near zero. The Taylor expansion at zero gives

$$\operatorname{erfi}(y/\sqrt{2}) = \sqrt{\frac{2}{\pi}}\left\{y + \frac{y^3}{6} + O(y^5)\right\}. \tag{79}$$

Therefore

$$\frac{m'(y)}{m(y)} = -\frac{2y}{3} + O(y^3), \qquad y \to 0. \tag{80}$$

Thus the score extends continuously to zero with value zero. Since $m(y) > 0$, it is continuous and bounded elsewhere. We have shown that its limit is zero in both tails. Together this shows it is bounded on $\mathbb{R}$. □

Proof of Proposition 9

In the following we suppress the coefficient subscripts. Write $T = \tau_g^2$ and $V = \varphi_{gl}$. Then

$$\begin{aligned} &T \sim \text{BetaPrime}(a_G, a_2), \qquad b \mid T, V \sim \text{Normal}(0, TV), \\ &V \sim \text{Beta}(c_g, (p_g - 1)c_g) \quad \text{if } p_g > 1, \end{aligned} \tag{81}$$

with T independent of V. For $p_g = 1$, we let $V = 1$.

We start with the behavior near zero. The mixture density is

$$p(b) = \int k_b(t, v)\, d(P_T \otimes P_V)(t, v), \tag{82}$$

where $k_b(t, v) = e^{-b^2/(2tv)}/\sqrt{2\pi tv}$. For any sequence $r_m \downarrow 0$ with $r_m > 0$, the kernels satisfy $k_{r_m}(t, v) \uparrow k_0(t, v)$ for every $t, v > 0$. Monotone convergence, followed by Tonelli's theorem and independence, gives

$$\begin{aligned} \lim_{m\to\infty} p(r_m) &= \int \lim_{m\to\infty} k_{r_m}(t, v)\, d(P_T \otimes P_V)(t, v) \\ &= \int \frac{1}{\sqrt{2\pi tv}}\, d(P_T \otimes P_V)(t, v) \\ &= \frac{1}{\sqrt{2\pi}} \mathbb{E}(T^{-1/2}) \mathbb{E}(V^{-1/2}) = p(0). \end{aligned} \tag{83}$$

Since $p(b)$ is even and nonincreasing in $|b|$, this also establishes the limit as $|b| \to 0$.

The first negative half moment is

$$\begin{aligned}\mathbb{E}(T^{-1/2}) &= \frac{1}{B(a_G, a_2)} \int_0^\infty t^{a_G - 3/2}(1+t)^{-a_G - a_2}\, dt \\ &= \begin{cases} \dfrac{B(a_G - 1/2, a_2 + 1/2)}{B(a_G, a_2)}, & a_G > 1/2, \\ +\infty, & a_G \le 1/2. \end{cases}\end{aligned} \tag{84}$$

When $p_g > 1$, the second moment is

$$\begin{aligned}\mathbb{E}(V^{-1/2}) &= \frac{1}{B(c_g, (p_g - 1)c_g)} \int_0^1 v^{c_g - 3/2}(1-v)^{(p_g - 1)c_g - 1}\, dv \\ &= \begin{cases} \dfrac{B(c_g - 1/2, (p_g - 1)c_g)}{B(c_g, (p_g - 1)c_g)}, & c_g > 1/2, \\ +\infty, & c_g \le 1/2. \end{cases}\end{aligned} \tag{85}$$

The finiteness conditions follow from the powers $t^{a_G - 3/2}$ and $v^{c_g - 3/2}$ at zero. Hence, $p(0) < \infty$ exactly when $a_G > 1/2$ and $c_g > 1/2$, and otherwise $p(b) \to \infty$ as $|b| \to 0$. For $p_g = 1$, $\mathbb{E}(V^{-1/2}) = 1$, so only $a_G > 1/2$ is needed.

We now show the tail behavior. For fixed $v > 0$, integrating over T gives

$$p(b \mid v) = \frac{1}{\sqrt{2\pi v}\, B(a_G, a_2)} \int_0^\infty t^{a_G - 3/2}(1+t)^{-a_G - a_2} e^{-b^2/(2vt)}\, dt. \tag{86}$$

For $b \neq 0$, substitute $u = b^2/(2vt)$ and then integrate over V. This combined with Tonelli yields

$$|b|^{2a_2+1} p(b) = K \int_{(0,1]} \int_0^\infty v^{a_2} e^{-u} u^{a_2 - 1/2} \left(1 + \frac{2vu}{b^2}\right)^{-a_G - a_2} du\, dP_V(v), \tag{87}$$

with $K = \frac{2^{a_2}}{\sqrt{\pi}\, B(a_G, a_2)}$.

Now, let (b_m) be any sequence of nonzero numbers with $|b_m| \to \infty$. Define

$$F_m(u,v) = v^{a_2} e^{-u} u^{a_2 - 1/2} \left(1 + \frac{2vu}{b_m^2}\right)^{-a_G - a_2}, \quad D(u,v) \;= v^{a_2} e^{-u} u^{a_2 - 1/2}. \tag{88}$$

For every $u > 0$ and $0 < v \le 1$, $F_m(u,v) \to D(u,v)$ and $0 \le F_m(u,v) \le D(u,v)$. The dominating function is integrable with respect to $du\, dP_V(v)$, since

$$\int_{(0,1]} \int_0^\infty D(u,v)\, du\, dP_V(v) = \Gamma(a_2 + 1/2)\mathbb{E}(V^{a_2}) \le \Gamma(a_2 + 1/2) < \infty. \tag{89}$$

Dominated convergence therefore gives

$$\begin{aligned}\lim_{m \to \infty} |b_m|^{2a_2+1} p(b_m) &= K \int_{(0,1]} \int_0^\infty \lim_{m \to \infty} F_m(u,v)\, du\, dP_V(v) \\ &= \frac{2^{a_2}\Gamma(a_2 + 1/2)}{\sqrt{\pi}\, B(a_G, a_2)} \mathbb{E}(V^{a_2}),\end{aligned} \tag{90}$$

where

$$\mathbb{E}(V^{a_2}) = \begin{cases} \dfrac{B(c_g + a_2, (p_g - 1)c_g)}{B(c_g, (p_g - 1)c_g)}, & p_g > 1, \\ 1, & p_g = 1. \end{cases} \tag{91}$$

Since the sequence was arbitrary, this proves the tail limit as $|b| \to \infty$. □

Proof of Proposition 10

For fixed group g, write $\Sigma_g = \tau_g^2 \operatorname{diag}(\varphi_{g1}, \ldots, \varphi_{gp_g})$. Then

$$b_g \mid \tau_g^2, \varphi_g \sim \text{Normal}_{p_g}(0, \Sigma_g). \tag{92}$$

We are interested in the marginal density $p(b_g \mid \varphi_g)$ obtained by integrating out τ_g^2. For notational convenience, set $t = \tau_g^2$. Since $\tau_g^2 \sim \text{BetaPrime}(a_G, a_2)$, we have

$$p(t) = \frac{1}{B(a_G, a_2)} t^{a_G - 1}(1 + t)^{-(a_G + a_2)}. \tag{93}$$

Therefore,

$$\begin{aligned} p(b_g \mid \varphi_g) &= \int_0^\infty p(b_g \mid t, \varphi_g) p(t)\, dt \\ &= \int_0^\infty (2\pi)^{-p_g/2} t^{-p_g/2} \left(\prod_{l=1}^{p_g} \varphi_{gl} \right)^{-1/2} \exp\left\{ -\frac{1}{2t} \sum_{l=1}^{p_g} \frac{b_{gl}^2}{\varphi_{gl}} \right\} \\ &\quad \times \frac{1}{B(a_G, a_2)} t^{a_G - 1}(1 + t)^{-(a_G + a_2)} dt. \end{aligned} \tag{94}$$

Collecting terms gives

$$p(b_g \mid \varphi_g) = K_g \int_0^\infty t^{a_G - p_g/2 - 1}(1 + t)^{-(a_G + a_2)} \exp\left\{ -\frac{z_g}{t} \right\} dt, \tag{95}$$

where $K_g = (2\pi)^{-p_g/2} / B(a_G, a_2) \left(\prod_{l=1}^{p_g} \varphi_{gl} \right)^{-1/2}$ and $z_g = 1/2 \sum_{l=1}^{p_g} b_{gl}^2 / \varphi_{gl}$. Now use the change of variables $u = 1/t$, we have

$$p(b_g \mid \varphi_g) = K_g \int_0^\infty u^{a_2 + p_g/2 - 1}(1 + u)^{-(a_G + a_2)} \exp\{-z_g u\} du. \tag{96}$$

Recall the integral representation of the confluent hypergeometric function of the second kind (Olver et al., 2010b),

$$U(\eta, \nu, z) = \frac{1}{\Gamma(\eta)} \int_0^\infty e^{-zu} u^{\eta - 1}(1 + u)^{\nu - \eta - 1} du. \tag{97}$$

Matching terms gives $\eta = a_2 + p_g/2$, $\nu = 1 + p_g/2 - a_G$. Hence,

$$p(b_g \mid \varphi_g) = \frac{(2\pi)^{-p_g/2}}{B(a_G, a_2)} \left(\prod_{l=1}^{p_g} \varphi_{gl} \right)^{-1/2} \Gamma\left(a_2 + \frac{p_g}{2} \right) U\left(a_2 + \frac{p_g}{2}, 1 + \frac{p_g}{2} - a_G, z_g \right). \tag{98}$$

We now study the behavior at the group origin. Conditional on fixed φ_g in the interior of the simplex, $b_g \to 0$ implies $z_g \to 0$. The small-argument behavior of $U(\eta, \nu, z)$ (Olver et al., 2010b) implies that, when $\nu > 1$,

$$U(\eta, \nu, z) \sim \frac{\Gamma(\nu - 1)}{\Gamma(\eta)} z^{1-\nu}, \qquad z \to 0. \tag{99}$$

Here $\nu > 1$ is equivalent to $1 + \frac{p_g}{2} - a_G > 1$ or $a_G < p_g/2$. Therefore, if $a_G < p_g/2$, the joint density diverges as $b_g \to 0$. When $a_G = p_g/2$, we have $\nu = 1$ and the divergence is logarithmic. When $a_G > p_g/2$, the density is bounded at the group origin (Olver et al., 2010b). This proves the result. □

Proof of Proposition 11

Fix a group g with $p_g \geq 2$ and two distinct coordinates j, k. We first study the covariance at a conditional level. Condition on $\tau_g^2 = t > 0$, and write $X = \varphi_{gj}$ and $Y = \varphi_{gk}$. Then

$$\kappa_{gj} = f_t(X), \qquad \kappa_{gk} = f_t(Y), \qquad f_t(x) = \frac{1}{1 + tx}. \tag{100}$$

Since the allocation vector is independent of the group scale, we can condition on t while leaving its Dirichlet law unchanged. Given $X = x$, the remaining allocations sum to $1 - x$, and the conditional Dirichlet distribution gives

$$Y \mid X = x \overset{d}{=} (1 - x)U, \tag{101}$$

where $U \sim \text{Beta}(c_g, (p_g - 2)c_g)$ if $p_g > 2$, and $U = 1$ if $p_g = 2$. The law of U does not depend on x, therefore

$$h_t(x) := \mathbb{E}(\kappa_{gk} \mid X = x, \tau_g^2 = t) = \mathbb{E}\left\{ \frac{1}{1 + t(1 - x)U} \right\}. \tag{102}$$

For $x < x'$ and $u > 0$,

$$\frac{1}{1 + t(1 - x)u} < \frac{1}{1 + t(1 - x')u}. \tag{103}$$

Since $U > 0$, h_t is strictly increasing, whereas f_t is strictly decreasing.

Let X' be an independent copy of $X \sim \text{Beta}(c_g, (p_g - 1)c_g)$. Conditioning first on X gives

$$\begin{aligned} \text{Cov}(\kappa_{gj}, \kappa_{gk} \mid \tau_g^2 = t) &= \mathbb{E}\{f_t(X)h_t(X)\} - \mathbb{E}\{f_t(X)\}\mathbb{E}\{h_t(X)\} \\ &= \frac{1}{2}\mathbb{E}\left\{[f_t(X) - f_t(X')][h_t(X) - h_t(X')]\right\}. \end{aligned} \tag{104}$$

The second equality follows by expanding the product and using independence for its cross terms. Since the functions are bounded, all expectations are finite. Whenever $X \neq X'$ strict monotonicity makes the product negative. Since $\Pr(X \neq X') = 1$, this proves the strict negative conditional covariance.

We now consider the unconditional covariance case. Assume $a_1 = Ga_G$ and $p_g c_g = a_G$. Proposition 2 gives the representation

$$\begin{aligned} \xi \sim \text{Gamma}(a_2, 1), \qquad & Z_{gj}, Z_{gk} \overset{\text{iid}}{\sim} \text{Gamma}(c_g, 1), \\ \lambda^2_{g\ell} = \frac{Z_{g\ell}}{\xi}, \qquad & \kappa_{g\ell} = \frac{\xi}{\xi + Z_{g\ell}}, \quad \ell \in \{j, k\}, \end{aligned} \tag{105}$$

where the gamma is parametrized in terms of shape and rate, and ξ, Z_{gj}, Z_{gk} are mutually independent.

Given ξ, the two shrinkage factors are independent with the same conditional mean. Writing $Z \sim \text{Gamma}(c_g, 1)$, define

$$m_g(x) = \mathbb{E}\left(\frac{x}{x + Z}\right), \qquad x > 0. \tag{106}$$

Thus $\mathbb{E}(\kappa_{gj} \mid \xi) = \mathbb{E}(\kappa_{gk} \mid \xi) = m_g(\xi)$. Therefore, the law of total covariance yields

$$\begin{aligned} \text{Cov}(\kappa_{gj}, \kappa_{gk}) &= \mathbb{E}\{\text{Cov}(\kappa_{gj}, \kappa_{gk} \mid \xi)\} + \text{Cov}\{\mathbb{E}(\kappa_{gj} \mid \xi), \mathbb{E}(\kappa_{gk} \mid \xi)\} \\ &= \text{Var}\{m_g(\xi)\}. \end{aligned} \tag{107}$$

To establish strict positivity, take $0 < x < x'$. Since $Z > 0$ almost surely,

$$m_g(x') - m_g(x) = \mathbb{E}\left\{\frac{Z(x' - x)}{(x + Z)(x' + Z)}\right\} > 0. \tag{108}$$

Thus m_g is strictly increasing. Since ξ is nondegenerate. $m_g(\xi)$ is nonconstant. Moreover it is bounded between zero and one. Consequently,

$$\text{Cov}(\kappa_{gj}, \kappa_{gk}) = \text{Var}\{m_g(\xi)\} > 0. \tag{109}$$

□

Shrinkage factors moments

For $p_g \geq 2$, the beta integral for φ_{gl} gives

$$\mathbb{E}(\kappa^m_{gl} \mid t) = \frac{1}{B(c_g, (p_g - 1)c_g)} \int_0^1 \frac{v^{c_g - 1}(1 - v)^{(p_g - 1)c_g - 1}}{(1 + tv)^m}\, dv, \tag{110}$$

which is the Euler representation of ${}_2F_1(m, c_g; p_g c_g; -t)$ (Olver et al., 2010b).

5.2 Additional simulations on $\mathbf{m_{eff,g}}$

We present in Figure 12 the prior predictive distributions for $\text{m}_{\text{eff},g}$ under the alternative that $a_2 = 0.1$, This configuration encourages fatter tails of the marginal prior, leading to a stronger tendency towards variable selection type behavior, where most panels show relatively more mass at the edges of the shrinkage coefficient space.

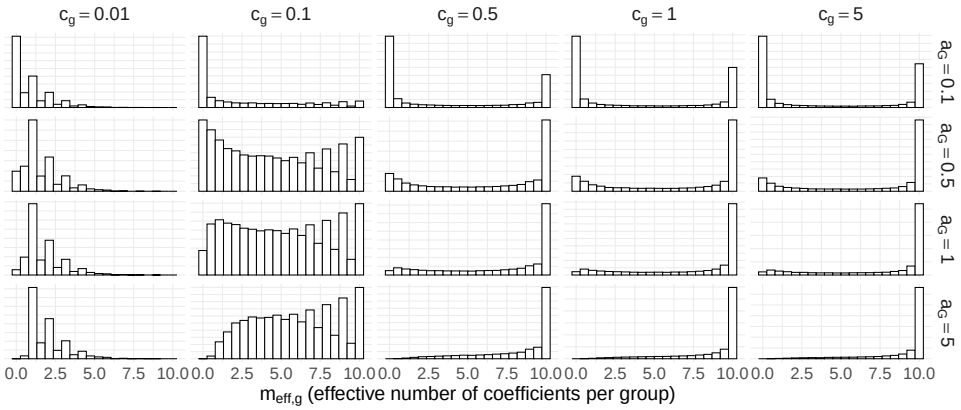


Figure 12: Prior predictive distributions of the group-wise effective number of nonzero coefficients ($m_{eff,g}$) under different combinations of a_G, c_g for groups of size 10. Rows vary a_G, the concentration parameter of the group-level Dirichlet prior, and columns vary c_g, the within-group concentration. We set $a_2 = 0.1$, $G = 10$, $p_g = 20, \forall g$.

5.3 Experiments

In the following we present additional results for the simulations.

Simulations: Lower-dimensional results

Figure 13 displays ROC curves for the various priors across the considered low-dimensional scenarios. Overall, Group-R2 priors tend to outperform their non-grouped counterparts, particularly in settings where the signal is concentrated within specific groups. For example, this advantage is clearly visible for the `R2-1.0` prior, which shows consistently better performance across all data-generating processes. However, as the concentration parameter a_G decreases the performance gap narrows. For instance, under the `R2-0.1` prior, the benefit of grouping becomes less pronounced.

Simulations: High-dimensional results

We show ΔRMSE in Figure 14. The results show that the Group-R2 priors consistently improve RMSE across distributed signal settings in the high-dimensional case. In contrast, no improvement is achieved under the Random Coefficients case. However performance does improve in this scenario as R^2 increases. Even though improvement is not strong in this case, including grouping structures in the prior is not detrimental either, as the distributions of ΔELPD are centered around zero. Concentrated signals are again challenging, however the loss involved by including groups is small.

We present the partitioned ΔRMSE results in Figures 15 and 16, corresponding to

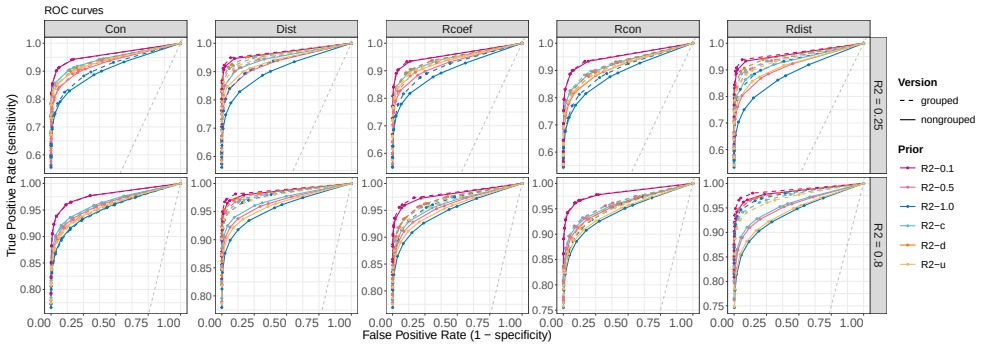


Figure 13: **ROC curves for the lower-dimensional scenario ($p = 100, n = 200$).** Grouped and non-grouped R^2 priors are compared. The grey line shows the diagonal of the unit square. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

truly zero and truly nonzero coefficients, respectively. The overall trends mirror those observed in the low-dimensional setting. Grouped versions of the R2 prior consistently outperform their non-grouped counterparts in detecting noise, with the exception of the concentrated signal scenario. In terms of signal recovery, certain priors lead to improved performance. As discussed earlier, the uniform R2 prior appears to offer a robust default across scenarios. However, we believe that incorporating well-informed prior knowledge about the grouping structure or sparsity pattern could lead to further gains.

Figure 17 shows the ROC curves for the high-dimensional case. The results again parallel the low-dimensional setting: under high noise and sparse signals, grouping structures help distinguish signal from noise more accurately on average, improving overall classification performance.

5.4 Comparison with further grouped shrinkage priors

The experiments of Section 3 isolate the effect of grouping by contrasting each Group-R^2 prior with its own nongrouped counterpart. This appendix instead compares the Group-R^2 prior against five grouped shrinkage priors proposed elsewhere in the literature. These are the Bayesian group lasso with spike and slab (BGL-SS) and the Bayesian sparse group selection with spike and slab (BSGS-SS) of Xu and Ghosh (2015), the group horseshoe (GHS) of Xu et al. (2016), the grouped regression with adaptive shrinkage priors (GRASP) of Tew et al. (2025), and the group-informed g-prior of Agarwal et al. (2025), which we fit with the authors' `GiVSA` sampler and refer to as the g-prior throughout.

Writing $\mathcal{Q}$ for a quantity of interest, we report $\Delta\mathcal{Q} = \mathcal{Q}(\text{Group-}R^2) - \mathcal{Q}(M_{\text{ext}})$, where M_{ext} is one of the competing group-priors. Because the two Group-R^2 configurations of

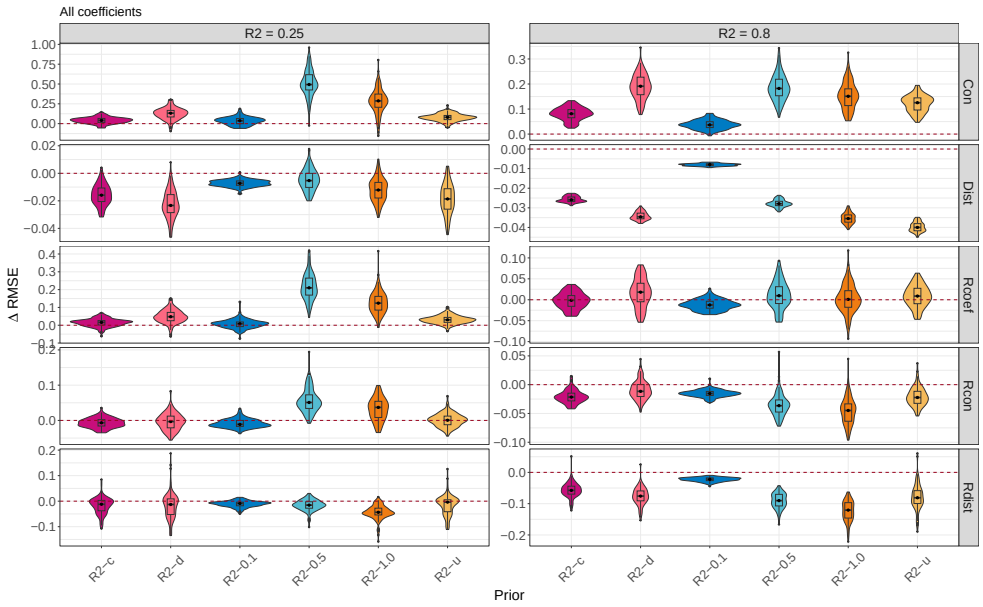


Figure 14: **ΔRMSE for the high-dimensional scenario ($p = 500$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Group-R2 priors improve parameter recovery across distributed and random signal settings, especially as a_G increases. In the concentrated case, no substantial improvement or deterioration is observed.

Section 3 encode opposite assumptions about how signal is spread within a group, we let the Group-R^2 side follow the data generating mechanism: `R2-c` ($a_G = 1$, $c_g = 0.5$) is used on the concentrated scenarios ("Con", "Rcon", "Rcoef") and `R2-d` ($a_G = 0.5$, $c_g = 1$) on the distributed ones ("Dist", "Rdist"). This gives each competing model the more favorable of the two matched alternatives in every panel. As before, $\Delta\text{ELPD} > 0$ and $\Delta\text{RMSE} < 0$ favor the Group-R^2 prior.

Competing models

Tables 2 and 3 collect the five priors in the notation of Section 2.1. All of them can be written as a scale mixture $b_{gl} \mid V_{gl} \sim \text{Normal}(0, V_{gl})$, and they differ in how the variance V_{gl} is built from a global, a group-level and a coefficient-level component, and in whether that construction admits an exact zero.

All five priors are fitted to a standardised design and a centered response, and are run as a single chain of 2000 draws after 2000 burn-in iterations, which matches the

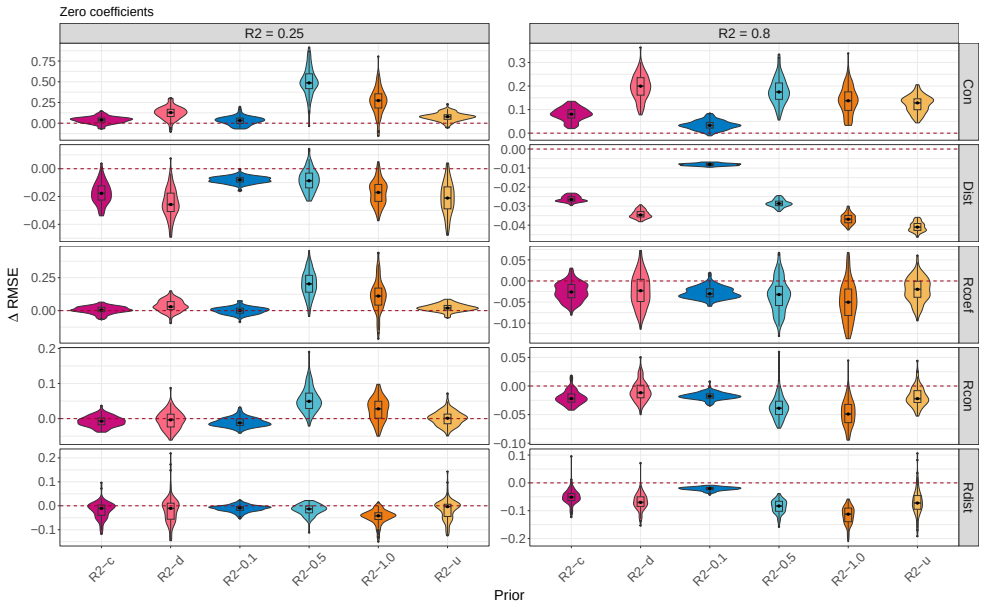


Figure 15: **ΔRMSE for truly zero coefficients for the high-dimensional scenario ($p = 500$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

4×1000 post-warmup draws used for the Group-R^2 models. Neither BGL-SS, BSGS-SS nor the g-prior included an intercept in their package implementations, so for these b_0 is drawn from its exact conditional $b_0 \mid b, \sigma^2 \sim \text{Normal}(\bar{y} - \bar{x}'b, \sigma^2/n)$ under a flat prior. For the g-prior, which samples over model space alone, b and σ^2 are drawn from the Normal–Inverse-Gamma conditional implied by the g-prior at each visited model, which is available in closed form. All coefficients are returned to the original scale of X before any comparison is made.

Functional differences

Three differences are relevant for what follows.

First, and most importantly, the competing models divide into *selection* priors and *continuous shrinkage* priors. BGL-SS, BSGS-SS and the g-prior place positive probability on $b_{gl} = 0$ exactly, either for a whole group or for individual coefficients, whereas GHS, GRASP and the Group-R^2 prior shrink continuously and set nothing exactly to zero. Since the data generating processes of Section 3 are genuinely sparse, we expect the selection priors to be more accurate on the truly zero coefficients, and the continuous

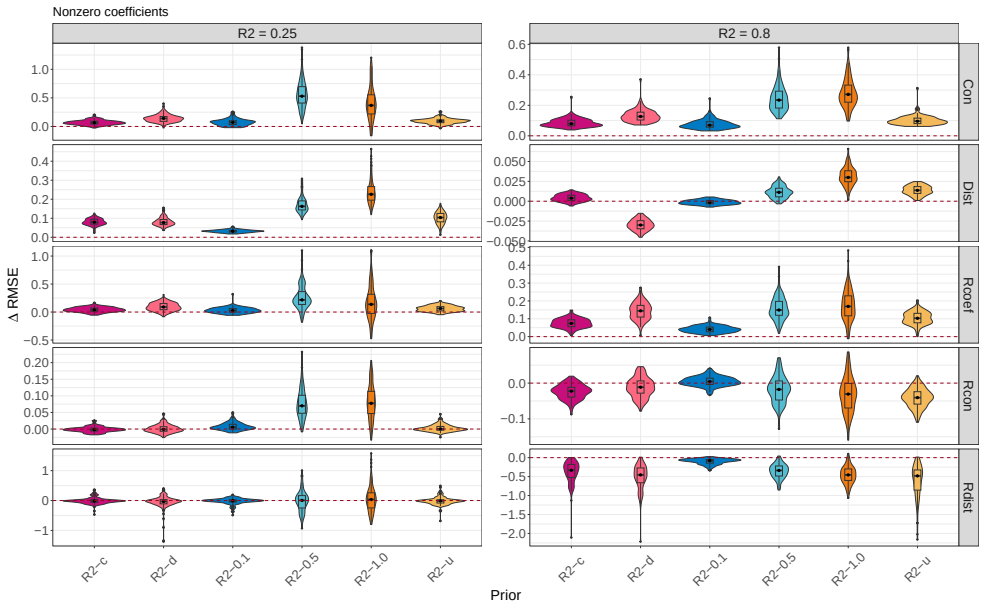


Figure 16: **ΔRMSE for truly nonzero coefficients for the high-dimensional scenario ($p = 500$, $n = 200$).** Values below the line indicate improvement in parameter recovery when considering groups. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

priors to pay less of a shrinkage penalty on the truly nonzero ones. Aggregate measures of parameter recovery are dominated by the zero coefficients and should therefore favor the selection priors.

Second, the priors differ in *where* the variance budget is controlled. The Group-R^2 prior is the only one that places an interpretable prior directly on R^2 and then partitions it, first across groups through $\phi \sim \text{Dirichlet}\, s(a_G)$ and then within groups through $\varphi_g \sim \text{Dirichlet}\, s(c_g)$. The simplex constraint means the groups compete for a fixed budget τ^2, so evidence for one group is evidence against the others. GHS and GRASP have a group scale δ_g^2 but no such constraint: their group scales are independent, and the total explained variance is whatever the product of the scales implies.

To anticipate the results, note that a symmetric Dirichlet $s(a_G)$ has mean $1/G$ in every coordinate, so an allocation in which every group carries a comparable share of the variance is close to the mean of the distribution, whereas an allocation in which a single group carries nearly all of it sits far in the tail, and increasingly so as G grows. We therefore expect the Group-R^2 prior to do comparatively well on the concentrated scenarios, where each of the G groups contains one signal, and comparatively less well

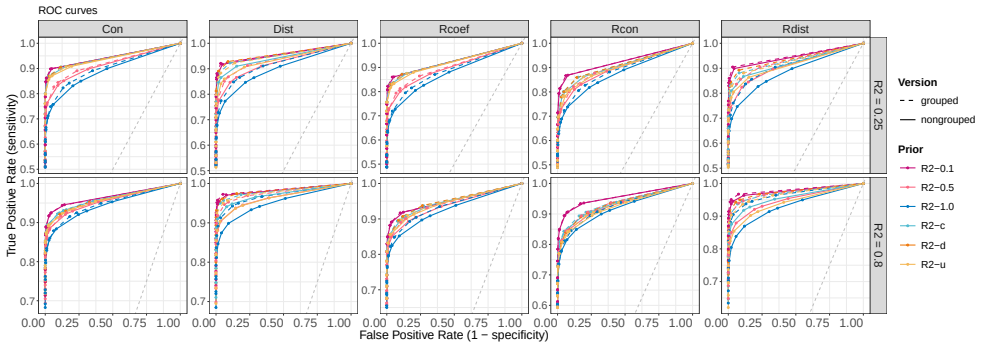


Figure 17: **ROC curves for the high-dimensional scenario ($p = 500, n = 200$).** Grouped and non-grouped R^2 priors are compared. The grey line shows the diagonal of the unit square. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

on the distributed ones, where a single active group must absorb the entire variance budget.

Third, the g-prior of Agarwal et al. (2025) is the only competing model whose informativeness grows with the design: $g = c_0 p^{2(1+\varsigma)}$ increases faster than p, so the prior is substantially more concentrated at $p = 500$ than at $p = 100$ and visits smaller models. Its two blocks should therefore not be read as one behavior observed twice.

Predictive performance

Figures 18 and 19 report ΔELPD for the two lower and higher-dimensional DGPs respectively. In the lower-dimensional setting the Group-R^2 prior is competitive but not dominant: it improves on BGL-SS (median ΔELPD $= 8.2$, favorable in 58% of replications), is indistinguishable from the g-prior (-1.6, 45%), and is outperformed by GHS (-6.8, 22%), GRASP (-7.3, 32%) and most clearly by BSGS-SS (-15.4, 1%). In the high-dimensional setting the ordering is preserved and the gaps widen: GHS (-14.4, 36%), the g-prior (-28.5, 36%) and BSGS-SS (-53.5, 3%) all predict better than the Group-R^2 prior.

Consistent with the budget argument above, the Group-R^2 prior is at its strongest on the concentrated scenarios, where it improves on every competing model except BSGS-SS at $p = 100$, and at its weakest on the distributed scenarios, where it is behind almost uniformly. Pooling over the competing models, the median ΔELPD is -1.1 on the concentrated scenarios against -11.4 on the distributed ones at $p = 100$, and -13.9 against -58.4 at $p = 500$. The widening of the gap with G is what the simplex argument predicts, since the distributed design requires one of G groups to hold essentially the whole budget and that configuration moves further into the tail of Dirichlet $s(a_G)$ as G

Table 2: Prior specification of the competing group-priors, in the notation of Section 2.1. Every prior is a scale mixture $b_{gl} \mid V_{gl} \sim \text{Normal}(0, V_{gl})$; δ_0 denotes a point mass at zero and C^+ a half-Cauchy distribution. Implementation details are in Table 3.

Prior	Coefficient variance V_{gl}	Hyperparameters
BGL-SS	A whole group is zero or not: $b_g \sim (1-\pi_0)\,\text{Normal}_{p_g}(0, \sigma^2\tau_g^2 I) + \pi_0\delta_0$	$\tau_g^2 \sim \text{Gamma}\left(\frac{p_g+1}{2}, \frac{\eta^2}{2}\right)$; η estimated by Monte Carlo EM
BSGS-SS	As BGL-SS, with a second spike and slab on each b_{gl} inside an active group	As BGL-SS
GHS	$\sigma^2\tau^2\delta_g^2\lambda_{gl}^2$	$\delta_g \sim \text{C}^+(0,1)$, $\lambda_{gl} \sim \text{C}^+(0,1)$
GRASP	$\sigma^2\tau^2\delta_g^2\lambda_{gl}^2$	$\frac{\delta_g^2}{1+\delta_g^2} \sim \text{Beta}(\alpha_D, \beta_D)$, $\frac{\lambda_{gl}^2}{1+\lambda_{gl}^2} \sim \text{Beta}(\alpha_G, \beta_G)$; all four shapes learned
g-prior (GiVSA)	$g\,\sigma^2 G_{\mathcal{M}}^{-1}$ for a selected model $\mathcal{M}$ and 0 otherwise, with $G_{\mathcal{M}} = \text{diag}(X_g' X_g)_{g\in\mathcal{M}}$	$g = c_0\, p^{2(1+\varsigma)}$ with $(c, c_0) = (0.01, 0.0025)$ and $\varsigma = 10^{-2}$

grows. The advantage of BSGS-SS is likewise concentrated where its extra within-group spike and slab is most useful.

The comparison also deteriorates as the signal strengthens. Moving from $R^2 = 0.25$ to $R^2 = 0.80$, the median ΔELPD falls against every competing model, from 18.6 to 2.7 (BGL-SS), 3.5 to -17.5 (g-prior), -2.5 to -15.8 (GHS), -3.2 to -13.2 (GRASP) and -9.2 to -23.8 (BSGS-SS) at $p = 100$, with the same pattern at $p = 500$. This is a property of the hyperparameters used here rather than of the prior class. Both Group-R^2 configurations in this comparison set $R^2 \sim \text{Beta}(1, 2)$, with prior mean $1/3$ and $\Pr(R^2 > 0.8) = 0.04$, so the prior is close to the truth when $R^2 = 0.25$ and places almost no mass near it when $R^2 = 0.80$, and τ^2 is shrunk towards too small a total variance. None of the competing priors constrains the total explained variance in this way. The scales of GHS and GRASP are independent and heavy-tailed, which allows better insulation in our simulations against high true signal strength. The same mechanism appears with the opposite sign in Section 3.3, where the rule $a_1 = Ga_G$ places the prior mean of R^2 well above the truth and the Group-R^2 prior is then the over-optimistic one. Predictive performance relative to these models therefore tracks how well the prior on R^2 is centered, which is the quantity the prior makes available for the user to set.

Two competing models fail numerically at $p = 500$ and are excluded from that block rather than reported. GRASP aborts 746 of 1000 fits: its adaptive update of the beta shape parameters fits a gamma proposal whose scale can become non-positive, after which the Metropolis acceptance ratio is not defined. BGL-SS completes every fit but returns a degenerate posterior in 86% of replications, with the residual scale overestimated by more than an order of magnitude. The resulting ΔELPD would be of order 10^5 and would reflect the breakdown of the sampler rather than the predictive

Table 3: Implementations used for the competing group-priors. Each prior is fitted with its authors' own code rather than a reimplementation. The last column records the step we add where the released code does not return posterior draws of all the quantities the evaluation of Section 3.1 requires.

Prior	Package	Sampler	Completion	Reference
BGL-SS[c]	`MBSGS`	Gibbs	σ^2 retained from the sweep in which b is stored, and b_0 drawn from its exact flat-prior conditional	Xu and Ghosh (2015)
BSGS-SS	`MBSGS`	Gibbs	As BGL-SS	Xu and Ghosh (2015)
GHS	`bayesreg`[a]	Gibbs	b_0 drawn from its exact flat-prior conditional	Xu et al. (2016)
GRASP[c]	`bayesreg`[a]	Gibbs with Metropolis steps for the shape parameters	As GHS	Tew et al. (2025)
g-prior (GiVSA)	`GiVSA`[b]	Population-based multiple-try Metropolis over model space	(b, σ^2) drawn from the closed-form Normal–Inverse-Gamma conditional of the g-prior given each visited model, then b_0 as above	Agarwal et al. (2025)

[a] The grouped fork of `bayesreg` released with Tew et al. (2025). The CRAN version has no group priors.
[b] The authors' own code. It samples over model space only and returns no draws of b or σ^2.
[c] GRASP is reported for $p = 100$ only, and BGL-SS is omitted from the $p = 500$ figures, as described in Section 5.4.

performance of the Group-R^2 prior. On the other hand, both behave well at $p = 100$. GHS shows a milder version of the same instability, with 5% of its $p = 500$ fits producing extreme values. These are retained and are visible as the upper outliers in Figure 19.

Parameter recovery

Figures 20 and 23 report ΔRMSE over all coefficients, and Figures 21–25 split it into the truly zero and the truly nonzero coefficients.

On aggregate the Group-R^2 prior recovers coefficients less accurately than every competing model, in both blocks, with median ΔRMSE between 0.12 and 0.57. The decomposition shows that this is driven entirely by the zero coefficients, where the median difference is positive throughout (0.15 to 0.66): the selection priors place mass at zero and the continuous Group-R^2 prior does not, so a coefficient that is truly zero is estimated exactly by the former and only shrunk by the latter. On the truly nonzero

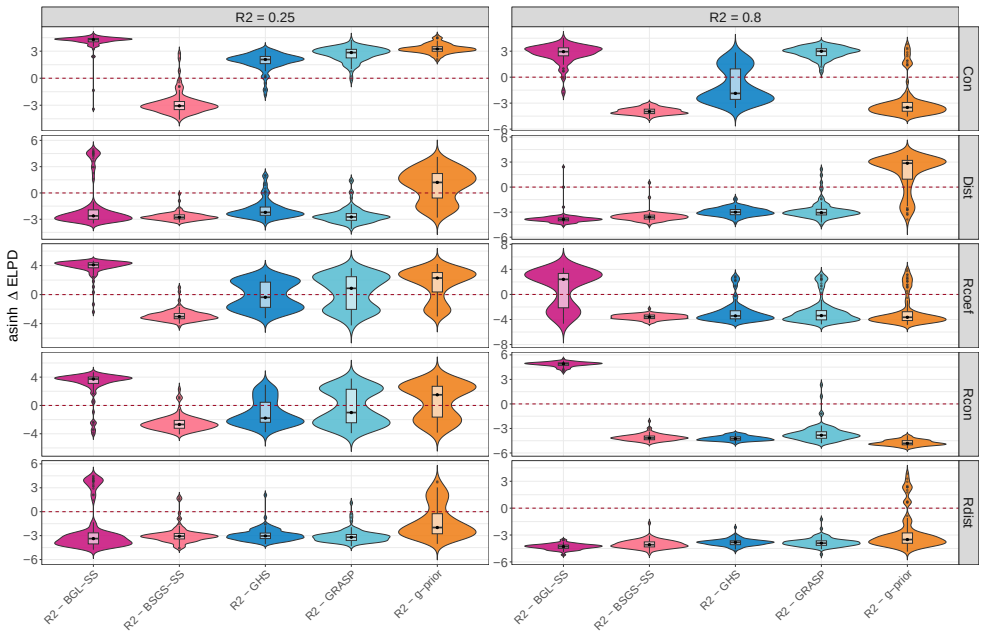


Figure 18: **ΔELPD (asinh-transformed) against the competing group-priors for the lower-dimensional scenario** ($p = 100$**,** $n = 200$**).** Values above the line indicate better predictive performance of the Group-R^2 prior. The Group-R^2 side is `R2-c` on the concentrated scenarios and `R2-d` on the distributed ones. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. The Group-R^2 prior is competitive with, but does not dominate, the competing group-priors.

coefficients the ordering reverses against four of the five competing models, with median differences of -0.11 (g-prior), -0.07 (BGL-SS), -0.06 (GRASP) and -0.04 (GHS) at $p = 100$: the Group-R^2 prior over-shrinks genuine signal less than they do. BSGS-SS is the exception and recovers the nonzero coefficients better as well ($+0.02$ at $p = 100$, $+0.11$ at $p = 500$), which is consistent with its within-group spike and slab being the only competing prior that can keep a group active while zeroing individual coefficients inside it.

This is the familiar trade-off between selection and continuous shrinkage rather than a deficiency specific to the Group-R^2 prior, and it has a direct practical implication: if the goal is to recover a sparse coefficient vector, a prior with an exact zero is preferable, and if the goal is to estimate the magnitude of the coefficients that are present, continuous shrinkage of the Group-R^2 form is competitive. Since the aggregate is dominated by the zeros in these designs, the aggregate alone gives a misleading picture of either.

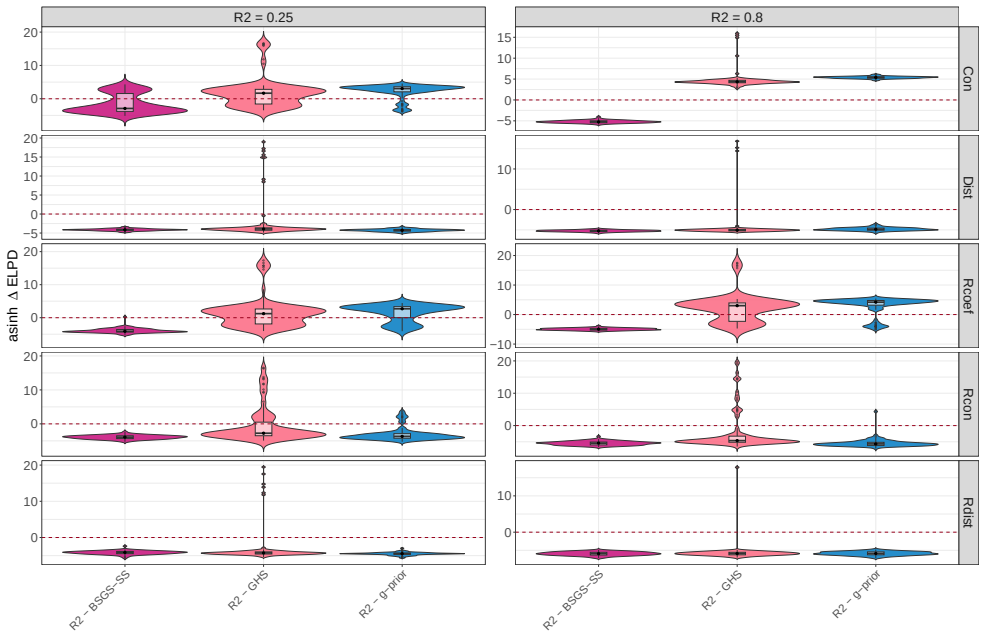


Figure 19: **ΔELPD (asinh-transformed) against the competing group-priors for the high-dimensional scenario** ($p = 500$, $n = 200$). Values above the line indicate better predictive performance of the Group-R^2 prior. GRASP and BGL-SS are omitted from this block because their samplers break down when $p > n$, as described in Section 5.4. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. The upper outliers in the GHS panels are replications in which its sampler becomes unstable.

Sampling diagnostics

Tables 4 and 5 report convergence diagnostics for the competing group-priors and for the Group-R^2 side of each panel, constructed as in Section 5.6. For every fit we take the largest $\widehat{R}$ and the smallest bulk effective sample size over the coefficients of the relevant type, and report the median of those quantities over the 100 replications. Two differences from Section 5.6 should be kept in mind. The five competing models are Gibbs samplers run as a single chain, so $\widehat{R}$ is computed by splitting that chain rather than across independent chains, and none of them produces the divergence diagnostics available for the Hamiltonian sampler used for the Group-R^2 models.

The Group-R^2 posterior is sampled more reliably than any of the competing models in every scenario and both dimensions, with $\widehat{R}$ at or below 1.004 and median effective sample sizes above 2000 and no divergent transitions. Among the competing models

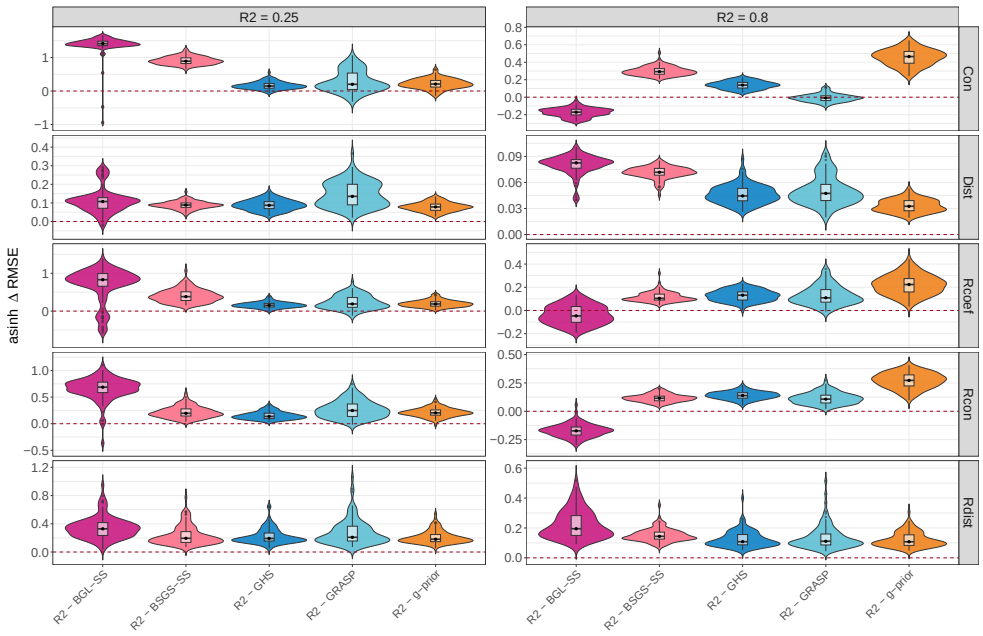


Figure 20: **ΔRMSE against the competing group-priors for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Aggregate recovery favors the competing models, and the decomposition in Figures 21 and 22 shows why.

the two spike and slab priors mix well, with $\widehat{R}$ below 1.011 throughout, while GHS, GRASP and the g-prior do not. At $p = 100$ their median $\widehat{R}$ reaches 1.40 and their median effective sample sizes fall to between 7 and 180 out of 2000 draws, and at $p = 500$ the g-prior returns effective sample sizes in the single digits. For the g-prior this is expected rather than anomalous, since it samples over model space and its marginal for a coefficient is a mixture of a point mass and a continuous component, which is strongly autocorrelated by construction. For GHS and GRASP it reflects the difficulty of their posterior geometry at these dimensions. The consequence for the comparison is that the deficits reported against GHS, GRASP and the g-prior in Section 5.4 rest on far fewer effective draws than the Group-R^2 side, and should be read as indicative rather than precise. The comparison against BGL-SS and BSGS-SS does not carry this caveat.

A further point concerns BGL-SS. In 27% of its fits at $p = 100$ and 6% at $p = 500$ at least one coefficient is exactly constant over the whole chain, so $\widehat{R}$ and the effective sample size are not defined for it. This is a property of a group spike and slab rather than a failure, since it means the group in question was never selected, and those coefficients

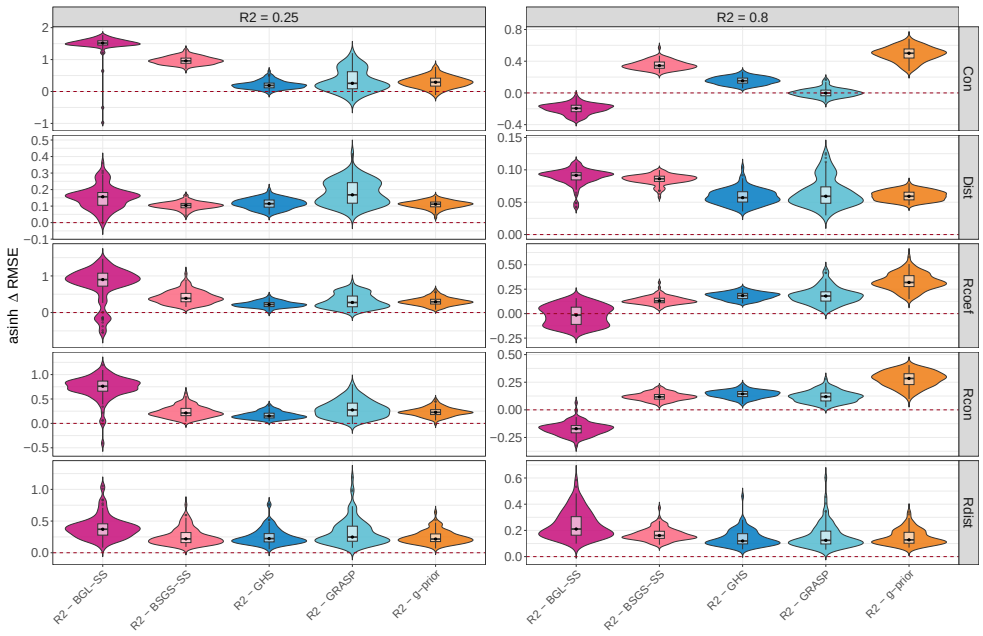


Figure 21: **ΔRMSE for truly zero coefficients against the competing group-priors ($p = 100$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. Priors that place mass at exactly zero recover the zero coefficients better.

are excluded from the medians reported below.

5.5 Calibration of the GIGG hyperparameters

This section gives the computational detail behind the calibration defined in (34) and reports the results relegated from Section 3.3.

Objective. The distribution of R^2_{GIGG} in (33) is available only by simulation. We draw τ^2, γ_g^2 and λ_{gl}^2 from the GIGG hierarchy and form V directly, leaving b marginalized rather than drawing it and using $\sum_{gl} b_{gl}^2$. While the two agree in distribution, the former removes a layer of Monte Carlo noise. In one dimension the optimal transport plan is the monotone rearrangement, so for two samples of equal size the objective in (34) reduces to a mean absolute difference of order statistics,

$$W_1 = \frac{1}{S}\sum_{s=1}^{S}\left|R^2_{(s)} - \tilde{R}^2_{(s)}\right|, \tag{111}$$

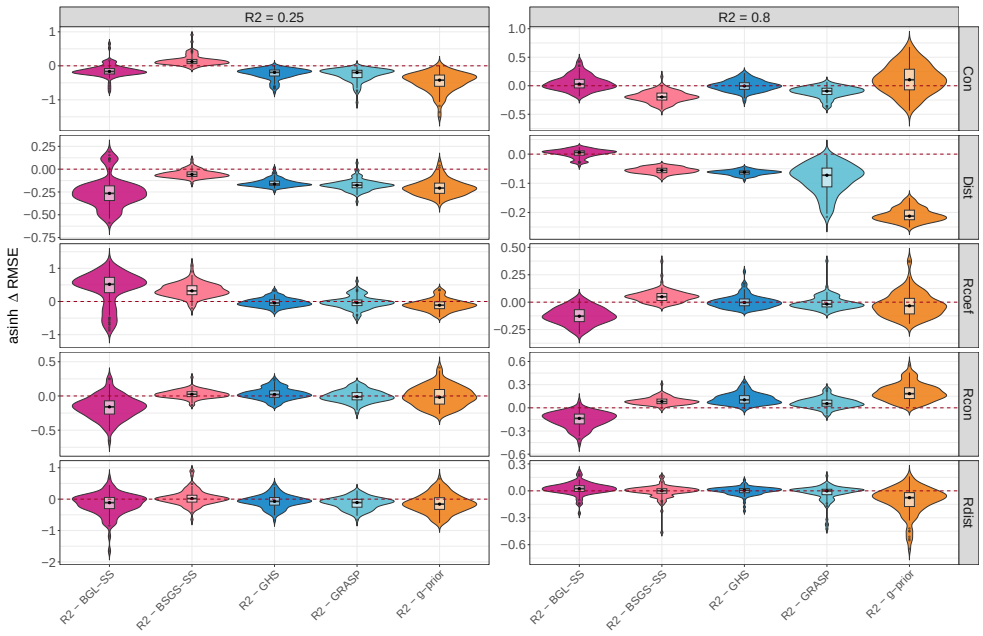


Figure 22: **ΔRMSE for truly nonzero coefficients against the competing group-priors ($p = 100$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting. The ordering of Figure 21 reverses against all competing models except BSGS-SS.

with $R^2_{(s)}$ the sorted draws from Equation (33) and $\tilde{R}^2_{(s)}$ the sorted draws from the target $\text{Beta}(Ga_G, a_2)$.

Optimization. Every candidate (a_g, b_g) is evaluated on the same fixed matrix of uniform random numbers, pushed through the relevant quantile functions. Using common random numbers in this way makes the objective a deterministic and smooth function of (a_g, b_g) rather than a noisy one, which is what allows a derivative free optimizer to converge. We minimize over the box $a_g \in [10^{-3}, 0.5]$ and $b_g \in [10^{-3}, 10]$. The upper bound on a_g keeps the pole of the marginal prior at the origin, which is the property that makes the GIGG a shrinkage prior, and the bound on b_g is a numerical guard. The search proceeds in two stages, a coarse grid over the box followed by a Nelder-Mead refinement from the best grid point, carried out on a logit reparameterization so that the constraints are respected without a penalty term. The grid stage is retained because the objective is flat over large parts of the box, since once b_g is small enough the implied R^2 is pinned near one whatever a_g does.

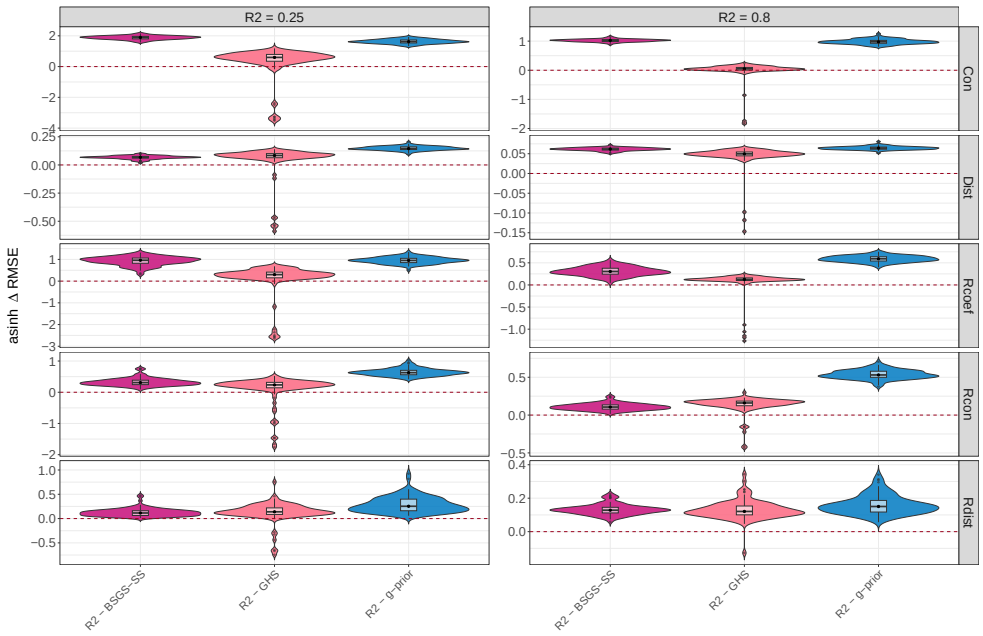


Figure 23: **ΔRMSE against the competing group-priors for the high-dimensional scenario ($p = 500$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. GRASP and BGL-SS are omitted from this block, as described in Section 5.4. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

Reported fit. The search uses $S = 4000$ draws. The distance reported in Table 1 is recomputed at the returned optimum on $S = 20{,}000$ independently seeded draws, so it is an out of sample measure of fit rather than the minimized value.

Figure 26 shows the resulting distributions. At $a_G = 0.1$ and $G = 10$ the target retains mass across the interior of the unit interval and the calibrated GIGG follows it reasonably closely. In every other panel both distributions concentrate near one, and the calibrated GIGG is the more concentrated of the two. This is the sense in which the small W_1 values of Table 1 record two distributions meeting at the boundary rather than agreeing in shape.

Relegated results. Figure 27 gives the predictive comparison for the high-dimensional setting, and Figures 28 and 29 give parameter recovery over all coefficients for both settings. The ordering across a_G is the same as in Figure 11 in every case.

Figures 30 to 33 split the same comparison by coefficient type, separating the

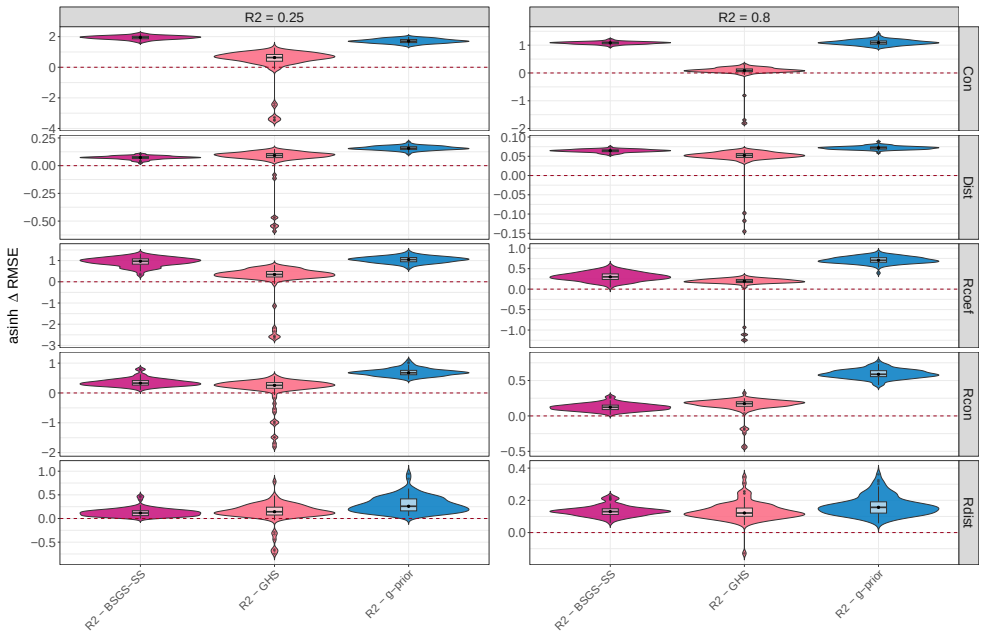


Figure 24: **ΔRMSE for truly zero coefficients against the competing group-priors ($p = 500$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

coefficients that are truly zero from those that are truly nonzero. The two types behave differently. On the truly zero coefficients the aggregate ordering is reproduced and the differences grow with both a_G and the dimension, with median ΔRMSE of 0.030 and 0.210 for $a_G = 0.5$ and $a_G = 1$ at $p = 100$ and 0.555 and 0.856 at $p = 500$. On the truly nonzero coefficients the corresponding medians are 0.026 and 0.025 at $p = 100$ and 0.039 and 0.104 at $p = 500$, so the difference stays small irrespective of a_G and of the dimension. Most of the aggregate difference in recovery is therefore attributable to the coefficients that are truly zero, which is consistent with the account of Section 3.3, since it is the mass that the implied prior on R^2 places away from zero that determines how strongly the null coefficients are shrunk. The exception is $a_G = 0.1$ at $p = 100$, where the Group-R2 prior recovers the zero coefficients better in 88% of replications, with a median ΔRMSE of -0.090, but the nonzero coefficients slightly worse, with a median of 0.046.

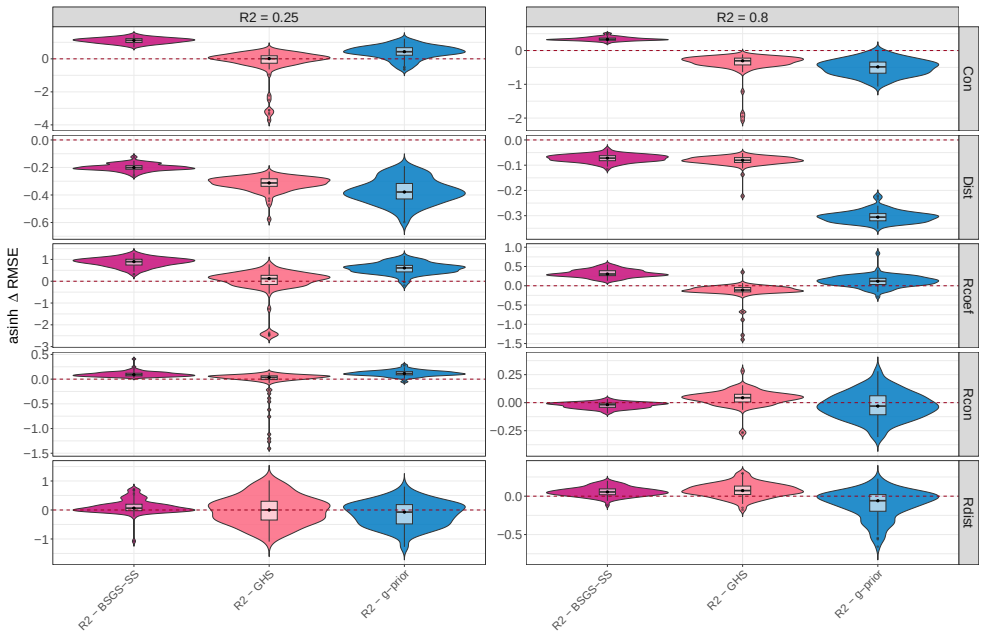


Figure 25: **ΔRMSE for truly nonzero coefficients against the competing group-priors ($p = 500$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R^2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

5.6 Sampling diagnostics for the GIGG comparison

Tables 6 and 7 report convergence diagnostics for the six models entering the comparison of Section 3.3, separately for the truly zero and the truly nonzero coefficients and for each data generating scenario. For every fit we take the largest $\widehat{R}$ and the smallest bulk effective sample size over the coefficients of the relevant type, and report the median of those quantities over the 100 replications. The last column gives the median number of divergent transitions per fit, out of the 4,000 post-warmup draws, which is a property of the fit rather than of a coefficient.

Three points stand out. First, the dependence on a_G runs in opposite directions for the two priors. Mixing in the GIGG deteriorates as a_G increases, whereas in the Group-R2 prior it is $a_G = 0.1$ that is the difficult case and the larger values that sample cleanly. Second, the deterioration in the GIGG is confined to the high-dimensional setting. At $p = 100$ all six models attain $\widehat{R}$ below 1.01 and effective sample sizes above 1,400, and the GIGG in fact mixes somewhat better than its Group-R2 counterpart. At $p = 500$ the picture changes, and `GIGG-1.0` reaches a median $\widehat{R}$ of up to 1.064 with

Table 4: Sampling diagnostics for the lower-dimensional scenario ($p = 100$, $n = 200$). Median over replications of the per-fit maximum $\widehat{R}$ and minimum bulk effective sample size, split by whether the coefficient is truly zero or truly nonzero, together with the median number of divergent transitions per fit. The Group-R^2 row is `R2-c` on the concentrated scenarios and `R2-d` on the distributed ones. Divergences are reported for the Group-R^2 models only, since the competing models are Gibbs samplers.

		$\widehat{R}$		ESS (bulk)		
Scenario	Model	zero	nonzero	zero	nonzero	Divergences
Con	Group-R^2	1.003	1.002	2,184	2,083	0
	BGL-SS	1.004	1.002	1,081	999	–
	BSGS-SS	1.006	1.003	775	638	–
	GHS	1.051	1.028	107	123	–
	GRASP	1.179	1.090	40	69	–
	g-prior	1.155	1.106	7	12	–
Dist	Group-R^2	1.003	1.002	2,375	3,435	0
	BGL-SS	1.003	1.002	1,170	1,625	–
	BSGS-SS	1.004	1.006	922	442	–
	GHS	1.101	1.012	176	187	–
	GRASP	1.397	1.068	117	37	–
	g-prior	1.066	1.051	24	25	–
Rcoef	Group-R^2	1.003	1.002	2,246	2,177	0
	BGL-SS	1.004	1.003	1,064	1,033	–
	BSGS-SS	1.006	1.005	786	543	–
	GHS	1.071	1.026	133	114	–
	GRASP	1.280	1.092	53	41	–
	g-prior	1.086	1.109	16	11	–
Rcon	Group-R^2	1.003	1.002	2,095	2,064	0
	BGL-SS	1.005	1.003	834	810	–
	BSGS-SS	1.009	1.005	487	460	–
	GHS	1.061	1.033	105	120	–
	GRASP	1.292	1.178	30	43	–
	g-prior	1.121	1.074	10	17	–
Rdist	Group-R^2	1.003	1.002	2,114	3,339	0
	BGL-SS	1.002	1.002	1,147	1,649	–
	BSGS-SS	1.004	1.006	956	457	–
	GHS	1.107	1.008	156	292	–
	GRASP	1.353	1.036	93	53	–
	g-prior	1.072	1.041	22	33	–

a median effective sample size near 120 on the zero coefficients, values at which the posterior cannot be regarded as adequately explored. Third, divergent transitions are not aligned with the $\widehat{R}$ and effective sample size ordering. `R2-`0.1 records the largest counts anywhere in the table, up to a median of 130 per fit, while its $\widehat{R}$ and effective sample sizes remain acceptable, and `GIGG-`1.0 at $p = 100$ records fewer than 10 per fit. The models with the weakest diagnostics are therefore `GIGG-`1.0 and `GIGG-`0.5 at $p = 500$, which are also the two configurations with the largest apparent predictive advantage over the Group-R2 prior in Section 3.3, and the differences reported there should be read with that in mind.

Table 5: Sampling diagnostics for the high-dimensional scenario ($p = 500$, $n = 200$). Entries are defined as in Table 4. BGL-SS and GRASP are omitted from this block, as described in Section 5.4.

		$\widehat{R}$		ESS (bulk)		
Scenario	Model	zero	nonzero	zero	nonzero	Divergences
Con	Group-R^2	1.004	1.003	2,655	2,965	0
	BSGS-SS	1.007	1.005	1,098	1,260	–
	GHS	1.162	1.115	38	79	–
	g-prior	1.349	1.132	3	7	–
Dist	Group-R^2	1.004	1.002	2,857	3,090	0
	BSGS-SS	1.006	1.009	568	416	–
	GHS	1.232	1.016	88	107	–
	g-prior	1.031	1.058	31	19	–
Rcoef	Group-R^2	1.004	1.003	2,616	2,628	0
	BSGS-SS	1.007	1.006	1,098	1,074	–
	GHS	1.160	1.120	45	48	–
	g-prior	1.263	1.211	3	4	–
Rcon	Group-R^2	1.004	1.003	2,180	2,008	0
	BSGS-SS	1.011	1.008	659	567	–
	GHS	1.167	1.124	29	37	–
	g-prior	1.231	1.117	4	7	–
Rdist	Group-R^2	1.004	1.002	2,400	2,893	0
	BSGS-SS	1.005	1.007	651	448	–
	GHS	1.217	1.009	92	248	–
	g-prior	1.027	1.024	41	39	–

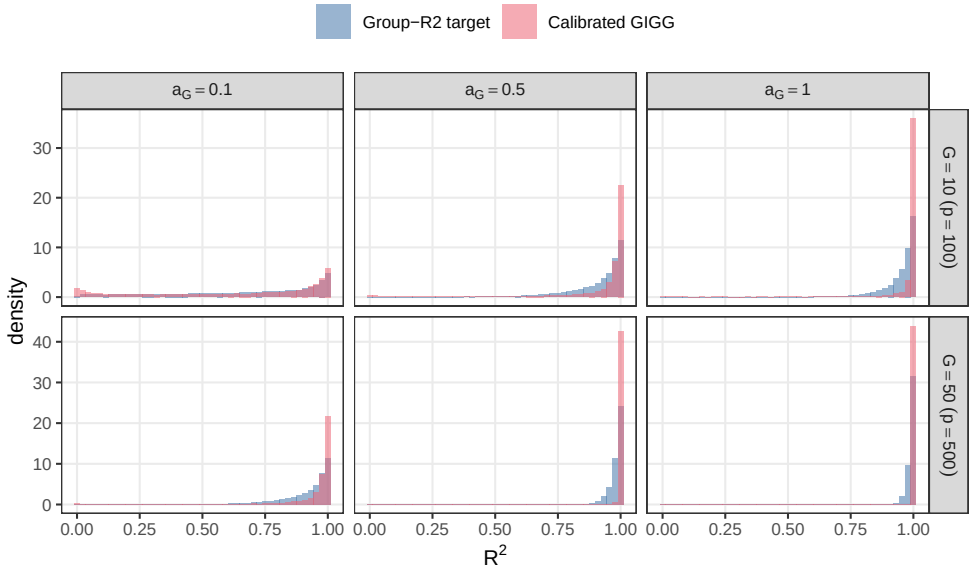


Figure 26: **Implied prior on R^2 under the calibrated GIGG against the Group-R2 target.** The target is the Beta(Ga_G, a_2) prior that the Group-R2 model places on R^2 directly. The calibrated GIGG uses the shapes of Table 1, which are the values fitted in the study. Both are drawn from 20,000 prior samples.

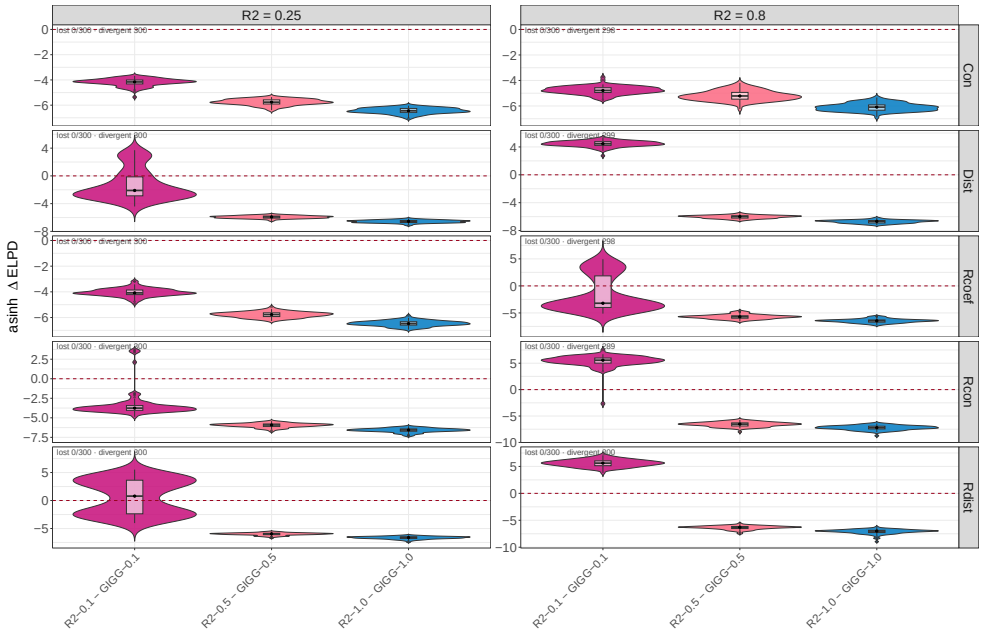


Figure 27: **ΔELPD (asinh-transformed) against the calibrated GIGG for the high-dimensional scenario** ($p = 500$, $n = 200$). Values above the line indicate better predictive performance of the Group-R2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

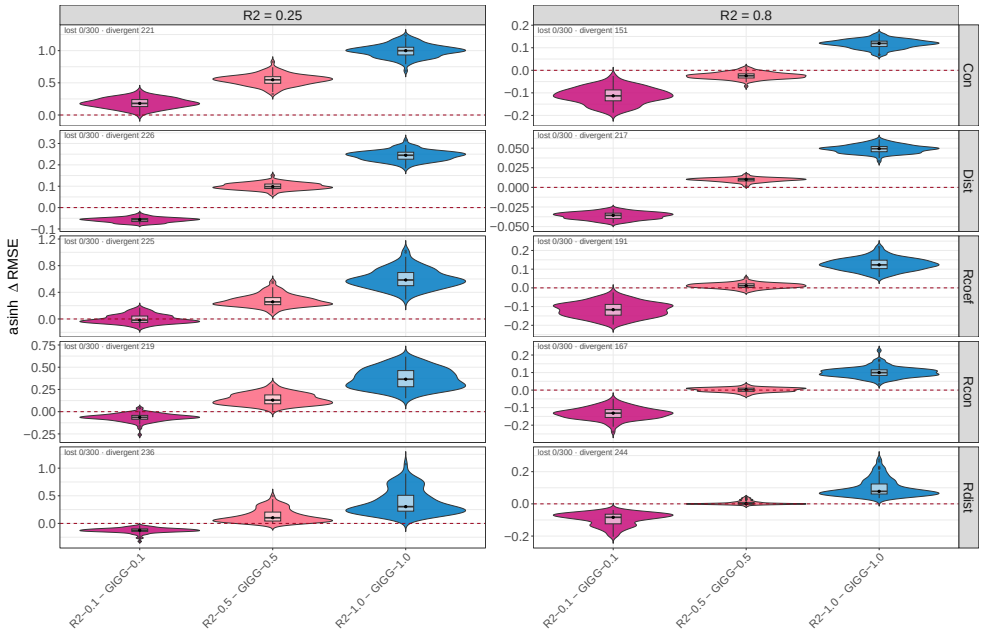


Figure 28: **ΔRMSE against the calibrated GIGG for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

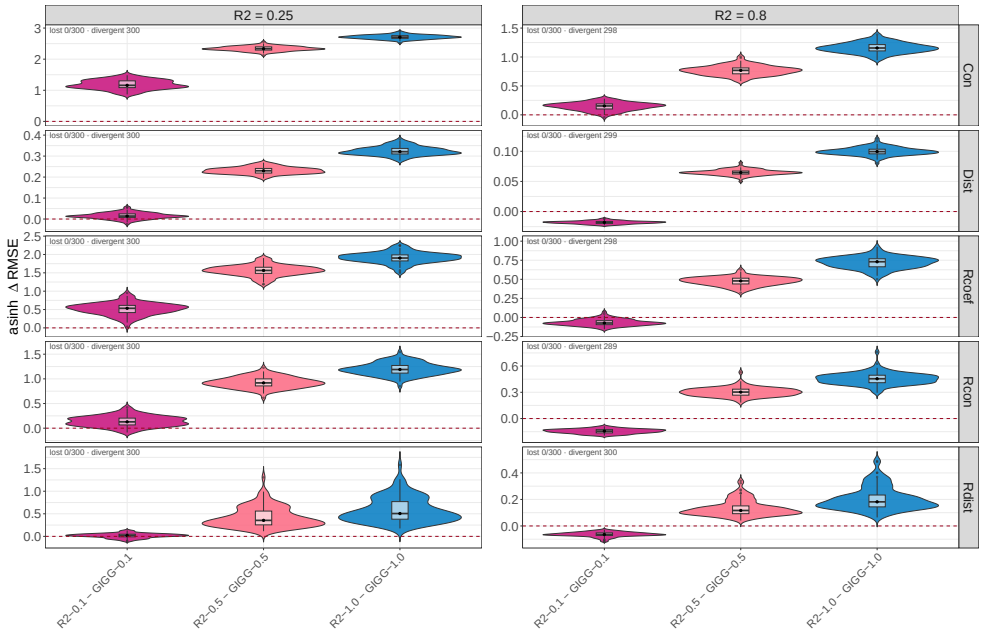


Figure 29: **ΔRMSE against the calibrated GIGG for the high-dimensional scenario ($p = 500$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

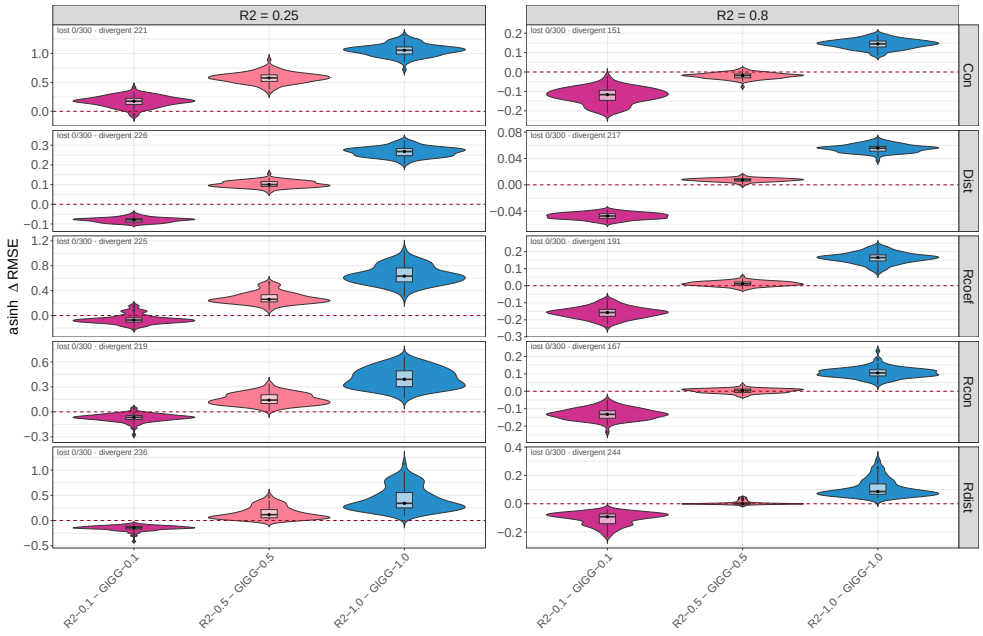


Figure 30: **ΔRMSE on the truly zero coefficients against the calibrated GIGG for the lower-dimensional scenario ($p = 100$, $n = 200$).** Values below the line indicate better parameter recovery of the Group-R2 prior. "Con" denotes the concentrated signal, "Dist" the distributed signal, "Rcon" and "Rdist" their random counterparts, and "Rcoef" the random coefficients setting.

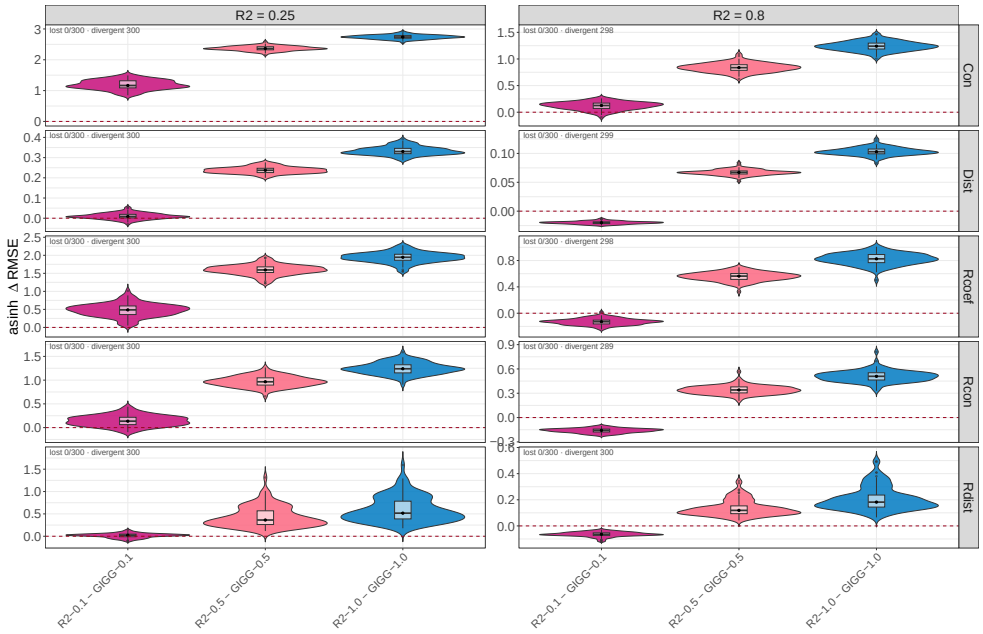


Figure 31: **ΔRMSE on the truly zero coefficients against the calibrated GIGG for the high-dimensional scenario ($p = 500$, $n = 200$).** Entries are defined as in Figure 30.

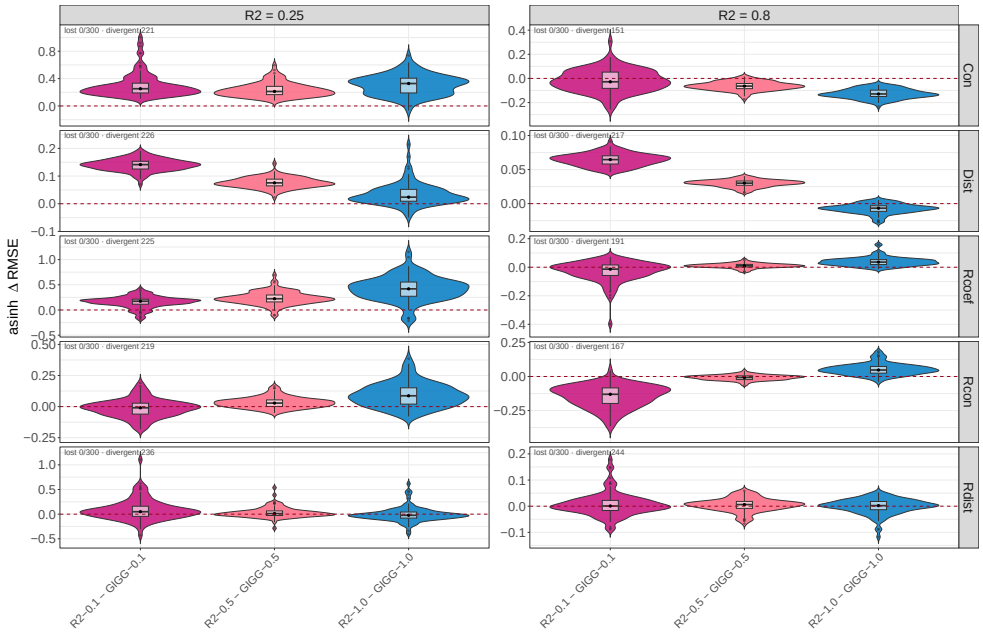


Figure 32: **ΔRMSE on the truly nonzero coefficients against the calibrated GIGG for the lower-dimensional scenario ($p = 100$, $n = 200$).** Entries are defined as in Figure 30.

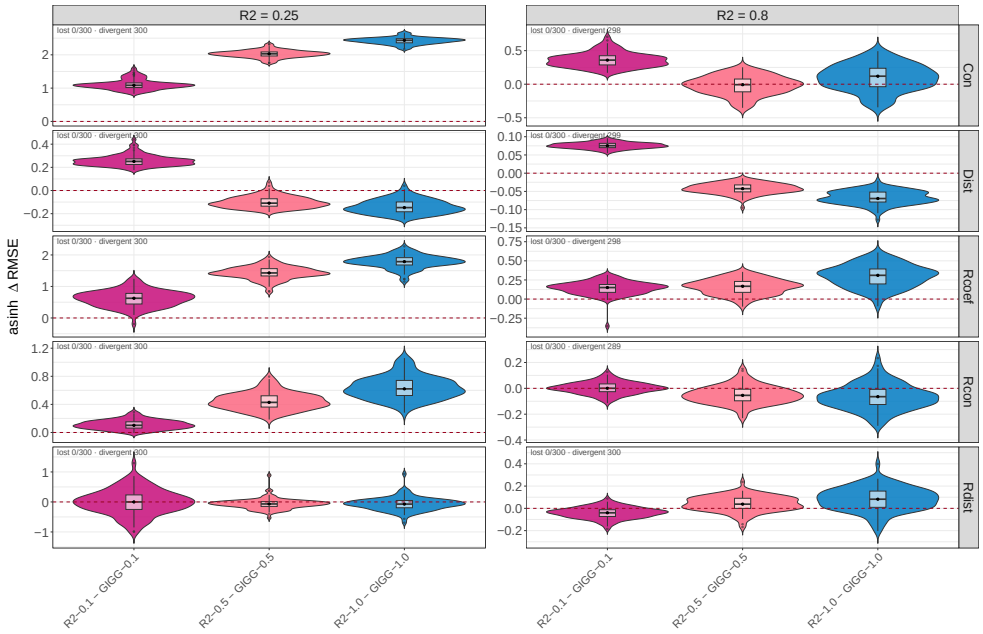


Figure 33: **ΔRMSE on the truly nonzero coefficients against the calibrated GIGG for the high-dimensional scenario ($p = 500$, $n = 200$).** Entries are defined as in Figure 30.

Table 6: Sampling diagnostics for the lower-dimensional scenario ($p = 100$, $n = 200$). Median over replications of the per-fit maximum $\widehat{R}$ and minimum bulk effective sample size, split by whether the coefficient is truly zero or truly nonzero, together with the median number of divergent transitions per fit out of 4,000 post-warmup draws.

		$\widehat{R}$		ESS (bulk)		
Scenario	Model	zero	nonzero	zero	nonzero	Divergences
Con	`R2-0.1`	1.007	1.004	1,492	1,406	14.5
	`GIGG-0.1`	1.004	1.002	2,824	2,659	0
	`R2-0.5`	1.003	1.002	1,913	1,856	0
	`GIGG-0.5`	1.003	1.002	2,439	2,429	0
	`R2-1.0`	1.003	1.002	1,968	2,031	0
	`GIGG-1.0`	1.003	1.002	1,909	1,959	9
Dist	`R2-0.1`	1.005	1.002	1,853	2,357	127
	`GIGG-0.1`	1.004	1.003	2,349	5,188	0
	`R2-0.5`	1.003	1.002	1,740	2,406	0
	`GIGG-0.5`	1.003	1.002	2,079	3,836	0
	`R2-1.0`	1.003	1.002	1,815	2,329	0
	`GIGG-1.0`	1.004	1.002	1,893	2,740	7
Rcoef	`R2-0.1`	1.005	1.003	1,760	1,919	113.5
	`GIGG-0.1`	1.004	1.003	2,626	3,446	0
	`R2-0.5`	1.003	1.002	1,830	2,109	0
	`GIGG-0.5`	1.003	1.002	2,252	2,868	0
	`R2-1.0`	1.003	1.002	1,921	2,153	0
	`GIGG-1.0`	1.003	1.002	1,807	2,055	8
Rcon	`R2-0.1`	1.006	1.004	1,530	1,553	44
	`GIGG-0.1`	1.004	1.002	2,310	2,471	0
	`R2-0.5`	1.003	1.002	1,800	1,891	0
	`GIGG-0.5`	1.003	1.002	2,204	2,486	0
	`R2-1.0`	1.003	1.002	1,907	2,009	0
	`GIGG-1.0`	1.003	1.002	1,747	1,909	10
Rdist	`R2-0.1`	1.005	1.002	1,707	2,301	129
	`GIGG-0.1`	1.004	1.002	2,106	4,008	0
	`R2-0.5`	1.004	1.002	1,544	2,415	0
	`GIGG-0.5`	1.004	1.002	1,833	3,437	0
	`R2-1.0`	1.003	1.002	1,800	2,344	0
	`GIGG-1.0`	1.004	1.002	1,700	2,532	9

Table 7: Sampling diagnostics for the high-dimensional scenario ($p = 500$, $n = 200$). Entries are defined as in Table 6.

		$\widehat{R}$		ESS (bulk)		
Scenario	Model	zero	nonzero	zero	nonzero	Divergences
Con	`R2-0.1`	1.009	1.006	875	1,333	117
	`GIGG-0.1`	1.005	1.003	2,903	3,322	0
	`R2-0.5`	1.005	1.003	1,439	1,863	0
	`GIGG-0.5`	1.016	1.012	536	814	33
	`R2-1.0`	1.004	1.003	1,556	2,001	0
	`GIGG-1.0`	1.064	1.045	117	221	73
Dist	`R2-0.1`	1.007	1.002	1,167	2,101	129.5
	`GIGG-0.1`	1.005	1.002	2,762	4,189	0
	`R2-0.5`	1.005	1.002	1,344	2,046	0
	`GIGG-0.5`	1.009	1.003	1,084	1,460	29
	`R2-1.0`	1.004	1.002	1,502	2,071	0
	`GIGG-1.0`	1.013	1.006	637	710	87
Rcoef	`R2-0.1`	1.008	1.008	1,002	1,136	119
	`GIGG-0.1`	1.004	1.004	2,666	2,971	0
	`R2-0.5`	1.005	1.004	1,428	1,675	0
	`GIGG-0.5`	1.011	1.011	679	742	32
	`R2-1.0`	1.004	1.003	1,558	1,766	0
	`GIGG-1.0`	1.049	1.049	149	188	70.5
Rcon	`R2-0.1`	1.010	1.007	822	860	122
	`GIGG-0.1`	1.005	1.003	2,171	2,263	0
	`R2-0.5`	1.005	1.004	1,152	1,293	0
	`GIGG-0.5`	1.016	1.013	516	486	33
	`R2-1.0`	1.004	1.003	1,326	1,471	0
	`GIGG-1.0`	1.053	1.044	171	149	71
Rdist	`R2-0.1`	1.008	1.002	1,038	2,128	129.5
	`GIGG-0.1`	1.005	1.002	2,226	3,273	0
	`R2-0.5`	1.005	1.002	1,263	2,083	0
	`GIGG-0.5`	1.009	1.003	957	1,487	33
	`R2-1.0`	1.004	1.002	1,365	1,971	0
	`GIGG-1.0`	1.018	1.007	522	672	81.5

References

Agarwal, S., Dutta, S., and Mukhopadhyay, M. (2025). "Bayesian Variable Selection Under High-Dimensional Settings with Grouped Covariates." *Bayesian Analysis*. URL https://doi.org/10.1214/25-BA1525 2, 7, 44, 48, 50

Aguilar, J. E. and Bürkner, P.-C. (2025). "Dependency-Aware Shrinkage Priors for High Dimensional Regression." *arXiv preprint arXiv:2505.10715*. 16

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711 – 1767.
URL https://doi.org/10.1214/23-EJS2136 1, 5

— (2025). "Generalized Decomposition Priors on R2." *Bayesian Analysis*, 1 – 34.
URL https://doi.org/10.1214/25-BA1524 1, 5, 9

Armagan, A., Dunson, D. B., Lee, J., Bajwa, W. U., and Strawn, N. (2013). "Posterior consistency in linear models under shrinkage priors." *Biometrika*, 100(4): 1011–1018.
URL https://doi.org/10.1093/biomet/ast028 12

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233.
URL https://doi.org/10.1080/02331888.2019.1662017 14

— (2021). "On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models." *Statistica Sinica*.
URL http://arxiv.org/abs/1807.06539 4, 11

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.
URL http://www.jstor.org/stable/2346101 19

Bentler, P. M. and Bonett, D. G. (1980). "Significance tests and goodness of fit in the analysis of covariance structures." *Psychological Bulletin*, 88(3): 588–606. Place: US Publisher: American Psychological Association. 1

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490.
URL https://doi.org/10.1080/01621459.2014.960967 5, 11

Boss, J., Datta, J., Wang, X., Park, S. K., Kang, J., and Mukherjee, B. (2023). "Group Inverse-Gamma Gamma Shrinkage for Sparse Linear Models with Block-Correlated Regressors." *Bayesian Analysis*, 1 – 30.
URL https://doi.org/10.1214/23-BA1371 2, 6, 13, 18, 20, 23, 26

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition. 18, 19

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Program-

ming Language." *Journal of Statistical Software*, 76(1): 1–32.
URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01 18

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.
URL http://www.jstor.org/stable/25734098 11, 12, 16, 37

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian Linear Regression with Sparse Priors." *The Annals of Statistics*, 43(5): 1986–2018.
URL https://www.jstor.org/stable/43818568 14

Castillo, I. and Vaart, A. v. d. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069 – 2101.
URL https://doi.org/10.1214/12-AOS1029 14

Coelho, C. A. and Alberto, R. P. (2021). "On the Distribution of the Product of Independent Beta Random Variables — Applications." In Ghosh, I., Balakrishnan, N., and Ng, H. K. T. (eds.), *Advances in Statistics - Theory and Applications: Honoring the Contributions of Barry C. Arnold in Statistical Science*, 85–117. Cham: Springer International Publishing.
URL https://doi.org/10.1007/978-3-030-62900-7_5 9

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515 – 534.
URL https://doi.org/10.1214/06-BA117A 6

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 1

Goodrich, B., Gabry, J., Ali, I., and Brilleman, S. (2020). "rstanarm: Bayesian applied regression modeling via Stan."
URL https://mc-stan.org/rstanarm 6

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15(1): 1593–1623. 18

Jeffrey, A., Zwillinger, D., Gradshteyn, I., and Ryzhik, I. (2007). "8–9 - Special Functions." In *Table of Integrals, Series, and Products (Seventh Edition)*, 859–1048. Boston: Academic Press, seventh edition.
URL https://www.sciencedirect.com/science/article/pii/B9780080471112500169 11, 15, 34

Kohns, D., Kallioinen, N., McLatchie, Y., and Vehtari, A. (2025). "The ARR2 Prior: Flexible Predictive Prior Definition for Bayesian Auto-Regressions." *Bayesian Analysis*, 1 – 32.
URL https://doi.org/10.1214/25-BA1512 1, 5

Kohns, D. and Potjagailo, G. (2025). "Flexible Bayesian MIDAS: time-variation, group-shrinkage and sparsity." *Journal of Business & Economic Statistics*, 1–17. 1

Kruijer, W., Rousseau, J., and Vaart, A. (2010). "Adaptive Bayesian Density Estimation with Location-Scale Mixtures." *Electronic Journal of Statistics*, 4. 8

Li, C. and Li, H. (2008). "Network-constrained regularization and variable selection for analysis of genomic data." *Bioinformatics*, 24(9): 1175–1182.
URL https://doi.org/10.1093/bioinformatics/btn081 1

Lin, J. (2016). "On The Dirichlet Distribution by Jiayu Lin." 5, 31

Miller, P. (2006). *Applied Asymptotic Analysis*. American Mathematical Society, 1 edition. 35

Mogliani, M. and Simoni, A. (2021). "Bayesian MIDAS penalized regressions: Estimation, selection, and prediction." *Journal of Econometrics*, 222(1): 833–860.
URL https://linkinghub.elsevier.com/retrieve/pii/S0304407620302207 1

Müller, P. and Quintana, F. A. (2004). "Nonparametric Bayesian Data Analysis." *Statistical Science*, 19(1): 95 – 110.
URL https://doi.org/10.1214/088342304000000017 3

Neal, R. M. (2011). *MCMC using Hamiltonian dynamics*.
URL http://arxiv.org/abs/1206.1901 18

Olver, F., of Standards, N. I., (U.S.), T., Lozier, D., Boisvert, R., and Clark, C. (2010a). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press.
URL https://books.google.de/books?id=3I15Ph1Qf38C 34, 35

Olver, F., Standards, N. I. o., (U.S.), T., Lozier, D., Boisvert, R., and Clark, C. (2010b). *NIST Handbook of Mathematical Functions Hardback and CD-ROM*. Cambridge University Press. 11, 33, 37, 40, 41, 42

Pas, S. L. v. d. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC. 11, 12

Pas, S. L. v. d., Kleijn, B. J. K., and Vaart, A. W. v. d. (2014). "The horseshoe estimator: Posterior concentration around nearly black vectors." *Electronic Journal of Statistics*, 8(2): 2585 – 2618.
URL https://doi.org/10.1214/14-EJS962 12, 14

Pas, S. L. v. d., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976 – 1000.
URL https://doi.org/10.1214/16-EJS1130 12, 14

Pas, S. L. v. d., Szabó, B., and Vaart, A. v. d. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221 – 1274.
URL https://doi.org/10.1214/17-BA1065 12

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018 –

5051.
URL https://doi.org/10.1214/17-EJS1337SI 12, 16

Polson, N. G. and Scott, J. G. (2012). "Local Shrinkage Rules, Lévy Processes and Regularized Regression." *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 74(2): 287–311.
URL https://doi.org/10.1111/j.1467-9868.2011.01015.x 2

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York. 19

Schmidt, D. F. and Makalic, E. (2020). "Log-Scale Shrinkage Priors and Adaptive Bayesian Global-Local Shrinkage Estimation."
URL https://arxiv.org/abs/1801.02321 5, 10

Schmüdgen, K. (2020). "Ten Lectures on the Moment Problem." ArXiv:2008.12698 [math].
URL http://arxiv.org/abs/2008.12698 9

Stan Development Team (2024). "Stan Modeling Language Users Guide and Reference Manual, Version 2.36."
URL http://mc-stan.org/ 18

Sun, H., Lin, W., Feng, R., and Li, H. (2014). "NETWORK-REGULARIZED HIGH-DIMENSIONAL COX REGRESSION FOR ANALYSIS OF GENOMIC DATA." *Statistica Sinica*, 24(3): 1433–1459. 1

Tew, S. Y., Schmidt, D. F., and Boley, M. (2025). "GRASP: Grouped Regression with Adaptive Shrinkage Priors."
URL https://arxiv.org/abs/2506.18092 2, 7, 10, 11, 44, 50

Vehtari, A., Gelman, A., and Gabry, J. (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432.
URL http://dx.doi.org/10.1007/s11222-016-9696-4 19

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667 – 718.
URL https://doi.org/10.1214/20-BA1221 19

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142 – 228.
URL https://doi.org/10.1214/12-SS102 19

Xu, X. and Ghosh, M. (2015). "Bayesian Variable Selection and Estimation for Group Lasso." *Bayesian Analysis*, 10(4).
URL https://projecteuclid.org/journals/bayesian-analysis/volume-10/issue-4/Bayesian-Variable-Selection-and-Estimation-for-Group-Lasso/10.1214/14-BA929.full 2, 7, 44, 50

Xu, Z., Schmidt, D. F., Makalic, E., Qian, G., and Hopper, J. L. (2016). "Bayesian

Grouped Horseshoe Regression with Application to Additive Models." In Kang, B. H. and Bai, Q. (eds.), *AI 2016: Advances in Artificial Intelligence*, Lecture Notes in Computer Science, 229–240. Cham: Springer International Publishing. 2, 7, 44, 50

Yanchenko, E., Bondell, H. D., and and, B. J. R. (2025a). "The R2D2 Prior for Generalized Linear Mixed Models." *The American Statistician*, 79(1): 40–49.
URL https://doi.org/10.1080/00031305.2024.2352010 1

Yanchenko, E., Irie, K., and Sugasawa, S. (2025b). "The Group R2D2 Shrinkage Prior for Sparse Linear Models with Grouped Covariates."
URL https://arxiv.org/abs/2412.15293 8

Zhang, Y. and Bondell, H. D. (2018). "Variable Selection via Penalized Credible Regions with Dirichlet–Laplace Global-Local Shrinkage Priors." *Bayesian Analysis*, 13(3): 823 – 844.
URL https://doi.org/10.1214/17-BA1076 14

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13.
URL https://doi.org/10.1080/01621459.2020.1825449 1, 5, 31

Article IV

Aguilar, J. E., and Bürkner, P.-C. (2025). Dependency-aware shrinkage priors for high-dimensional regression.

Preprint. Manuscript submitted for publication.

Available at https://arxiv.org/abs/2505.10715

Dependency-Aware Shrinkage Priors for High Dimensional Regression

Javier Enrique Aguilar[1,*] **and Paul-Christian Bürkner**[1,†]

[1] *Department of Statistics, TU Dortmund University, Germany*

Abstract: In high dimensional regression, global local shrinkage priors have gained significant traction for their ability to yield sparse estimates, improve parameter recovery, and support accurate predictive modeling. While recent work has explored increasingly flexible shrinkage prior structures, the role of explicitly modeling dependencies among coefficients remains largely unexplored. In this paper, we investigate whether incorporating such structures into traditional shrinkage priors improves their performance. We introduce dependency-aware shrinkage priors, an extension of continuous shrinkage priors that integrates correlation structures inspired by Zellner's g prior approach. We provide theoretical insights into how dependence alters the prior and posterior structure, and evaluate the method empirically through simulations and real data. We find that modeling dependence can improve parameter recovery when predictors are strongly correlated, but offers only modest gains in predictive accuracy. These findings suggest that prior dependence should be used selectively and guided by the specific inferential goals of the analysis.

Keywords and phrases: Prior specification, shrinkage priors, structured shrinkage, high dimensional regression, regularization.

1. Introduction

Regression analysis has long been a cornerstone of statistical modeling (Hastie et al., 2009; Gelman et al., 2020). As datasets have grown in size and complexity, the need for models capable of handling high dimensional settings while avoiding overfitting has become increasingly evident (Giraud, 2014; Wainwright, 2019). To address this challenge, a wide array of priors with strong theoretical and empirical properties has been proposed over the past decade (Carvalho et al., 2010; Brown and Griffin, 2010; Bhattacharya et al., 2015; Piironen and Vehtari, 2017; Ročková and George, 2018; Zhang et al., 2020).

Let $y_i \in \mathbb{R}$ denote the ith response variable, related to p explanatory variables $x_i' = (x_{i1}, \ldots, x_{ip}) \in \mathbb{R}^p$ through the linear regression model

$$y_i = \alpha + x_i' b + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\alpha \in \mathbb{R}$ is the intercept, $b = (b_1, \ldots, b_p)' \in \mathbb{R}^p$ is the vector of regression coefficients, and ε_i is the error term, assumed to satisfy $\varepsilon_i \sim \mathcal{N}(0, \sigma^2)$ with unknown residual variance σ^2 (Gelman et al., 2013).

*Corresponding author: e-mail: javier.aguilarr@icloud.com url: https://jear2412.github.io

†e-mail: paul.buerkner@gmail.com url: https://paul-buerkner.github.io

Although model (1) is standard, we are particularly interested in the high dimensional regime where $p > n$ (Tibshirani, 1996; Giraud, 2014). In this setting, a central challenge is the specification of prior distributions for the coefficients b and the residual variance σ^2 (Gelman, 2006; Tadesse and Vannucci, 2021). Introducing additional assumptions about the data-generating process can simplify this task. In particular, if we assume that the true model is sparse—that is, only a small subset of the coefficients are nonzero—then shrinkage priors become an attractive modeling choice (Johnstone and Silverman, 2004; Castillo and Vaart, 2012; Pas et al., 2016, 2017). Sparse solutions are heavily recommended: dense models not only hinder interpretability and statistical parsimony, but may also lead to computational inefficiencies and overfitting (Bühlmann and Van De Geer, 2011; Giraud, 2014; Hastie et al., 2015).

Among shrinkage priors, the class of continuous global-local shrinkage priors has gained significant attention (Mitchell and Beauchamp, 1988; Carvalho et al., 2010; Bhattacharya et al., 2015; Pas et al., 2016; Bhadra et al., 2016; Zhang et al., 2020; Tadesse and Vannucci, 2021). These priors balance sparsity and flexibility by combining global shrinkage, which controls the overall level of shrinkage, with local shrinkage, which allows individual coefficients to escape strong penalization. Their popularity stems from a combination of excellent theoretical properties, such as minimax optimality and desirable posterior concentration rates, and strong empirical performance (Castillo and Vaart, 2012; Castillo et al., 2015; Pas et al., 2016; Ghosh and Chakrabarti, 2017; Pas, 2021; Tadesse and Vannucci, 2021). Furthermore, their compatibility with efficient computational strategies and probabilistic programming frameworks has contributed to their widespread adoption (George and McCulloch, 1993; Piironen and Vehtari, 2017; Bhattacharya et al., 2016; Aguilar and Bürkner, 2023).

Incorporating prior knowledge about regression coefficients can, in principle, improve both parameter estimation and predictive performance, provided that the prior sufficiently reflects true underlying structure (Pas et al., 2017; Simpson et al., 2017). For instance, modeling multivariate dependencies among coefficients allows to capture joint uncertainty in related parameters (Gelman et al., 2013, 2020). While such approaches have been explored in non-shrinkage contexts, analogous developments within shrinkage prior frameworks remain relatively limited, since conditional independence of the coefficients is typically assumed (Zellner, 1996; Casella et al., 2010; Griffin and Hoff, 2024).

A foundational example of priors that incorporate dependencies is Zellner's g prior, which uses the design matrix to encode the dependence structure of the regression coefficients via their prior covariance matrix (Zellner, 1986; Maruyama and George, 2011; Li and Clyde, 2018). Although not originally conceived as a shrinkage prior, the g prior exhibits a global shrinkage structure and remains widely used in Bayesian model selection and averaging due to its closed-form marginal likelihood (Robert and Casella, 2004; Liang et al., 2008; Maruyama and George, 2011; Li and Clyde, 2018).

Some approaches attempt to encode coefficient dependence by treating regularized objectives as priors; specifically, by normalizing expressions of the form $b'\Omega b + R(b)$, where Ω is a positive semidefinite matrix and $R(b)$ is a penalty term

(Casella et al., 2010; Pauger and Wagner, 2019; Goplerud, 2021). Choosing $R(b)$ as a convex penalty leads to computationally tractable optimization problems (Tibshirani, 1996; Hastie et al., 2015). However, translating these penalties into fully Bayesian priors is not straightforward: the role of hyperparameters and their interaction with the likelihood can be difficult to interpret or tune (Casella et al., 2010; Hahn and Carvalho, 2015; Griffin and Hoff, 2024). Moreover, the key advantages of regularization methods, namely computational efficiency and the ability to produce sparse solutions, are often diminished in the Bayesian setting (Hahn and Carvalho, 2015).

Additionally, although these constructions superficially resemble frequentist regularization techniques (such as interpreting the log-prior as a penalty), the comparison can be misleading. The theoretical properties of Bayesian shrinkage priors and frequentist regularizers differ in fundamental ways (Simpson et al., 2017; Castillo, 2024). For example, under suitable conditions on the design matrix, the Lasso estimator can consistently recover sparse signals (Zhao and Yu, 2006). In contrast, the Laplace prior, despite its formal similarity to the Lasso penalty, tends to over-shrink coefficients due to its insufficient tail mass, often leading to biased estimates and poor uncertainty quantification (Castillo et al., 2015; Pas et al., 2017; Bhadra et al., 2019; Castillo, 2024). While initial analogies between Bayesian and frequentist approaches may be instructive, they can obscure important differences in parameter recovery, shrinkage behavior, and predictive performance.

A more recent and notable development is the Structured Shrinkage Prior proposed by Griffin and Hoff (2024), which incorporates dependence by constructing a covariance matrix as the elementwise product of the second-moment matrix of the local scales and an arbitrary structure-imposing matrix. While this framework generalizes existing approaches and offers conceptual elegance, it is strongly limited to shrinkage priors with unit second moments for the local scales. This constraint is nontrivial, as the second moment is sensitive to the choice of hyperparameters—an important tuning mechanism in high dimensional models—and may not equal one in practice. See Section 2 for further details.

While prior dependence structures have been explored in low dimensional asymptotics, their role in high dimensional models remains poorly understood. For instance, Hagar and Stevens (2024) show that in low dimensions, complex dependence encoded through copula-based priors may be overridden by the likelihood as sample size increases. However, their results rely on the Bernstein–von Mises theorem and do not apply in high dimensional settings, where the behavior of posterior dependence structures remains largely an open question (Ghosal et al., 2000; Ghosh, 2003; Castillo et al., 2015).

1.1. Contributions of our paper

The main aim of this paper is to investigate whether traditional continuous global-local shrinkage priors can benefit from the inclusion of dependency struc-

tures in high dimensional sparse settings, and to determine the conditions under which such benefits arise. To this end, we introduce a generalization of global-local shrinkage priors, which we term dependency-aware shrinkage priors (DASP). Our approach extracts correlation information from the design matrix of the predictors and combines it with the shrinkage scales in a data driven manner.

To evaluate the impact of incorporating dependence structures, we present both theoretical and empirical results. We characterize how introducing dependence modifies the structure of the prior and posterior distributions. We conduct simulation studies using both the true correlation matrix and our data-driven approach. The results show that incorporating dependence can improve parameter recovery when strong correlation exists among groups of coefficients. However, the gains in predictive performance are generally modest. We further validate our method on multiple real-world datasets, where we observe consistent patterns: dependency-aware priors may aid estimation in structured settings but do not yield substantial improvements in prediction. We conclude that, while prior dependence structures can be useful in specific inferential contexts, they are not universally beneficial and should not be used by default.

2. Methods

2.1. Shrinkage priors

Continuous global-local shrinkage priors are constructed as scale mixtures of normal distributions (West, 1987), yielding a hierarchical model for the regression coefficients. Each coefficient b_i is modeled as

$$b_i \mid \sigma, \tau, \lambda_i \sim \mathcal{N}(0, \sigma^2\tau^2\lambda_i^2), \quad \lambda_i \sim p(\lambda_i), \quad \tau \sim p(\tau), \quad \sigma \sim p(\sigma), \tag{2}$$

where τ is a global scale parameter that controls the overall level of shrinkage, and λ_i are local scale parameters that allow coefficient-specific deviations. The global scale reflects the belief that most coefficients are near zero, while the local scales allow signals to escape shrinkage when warranted by the data. This hierarchy encourages sharing of information about sparsity across coefficients while retaining flexibility at the individual level. We include the residual variance σ^2 within the prior variance of b_i, following standard practice, as this improves adaptivity to varying signal-to-noise ratios (Pas, 2021).

With appropriate choices for the priors on λ_i and τ, global-local shrinkage priors induce approximate sparsity, shrinking irrelevant signals strongly toward zero while avoiding excessive shrinkage of relevant signals. The combination of local and global scales also endows these priors with desirable theoretical properties, such as minimax optimality, posterior consistency, and optimal contraction rates under sparsity (Castillo and Vaart, 2012; Ghosh and Chakrabarti, 2017; Pas et al., 2016; Pas, 2021). Although they do not produce exact zeros, which may be required in some applications, model selection can be handled post

hoc by decoupling inference from selection (Hahn and Carvalho, 2015; Piironen et al., 2020).

The specific choices for the priors on λ_i and τ give rise to the wide array of shrinkage priors used today. Common examples include the Horseshoe, Three-Parameter Beta, Normal-Gamma, and Beta Prime priors, among others (Carvalho et al., 2010; Armagan et al., 2011; Brown and Griffin, 2010; Bai and Ghosh, 2019). These models belong to the family of shrinkage priors that treat the local scales λ_i as conditionally independent. More recently, alternative approaches have been proposed that introduce joint modeling of the local scales via a multivariate distribution; examples of this include the Dirichlet–Laplace, R2D2, and Generalized R2 Decomposition priors (Bhattacharya et al., 2015; Zhang et al., 2020; Aguilar and Bürkner, 2025).

2.2. Dependency-aware shrinkage priors

Regression coefficients are typically deemed as conditionally independent a priori under the shrinkage prior setup (2). We propose a natural extension of this model by considering dependencies in the regression coefficients via correlation matrix Ω. The model we propose is the following:

$$b \mid \sigma, \tau, \lambda \sim \mathcal{N}\left(0, \sigma^2\tau^2 \mathrm{D}_\lambda \,\Omega\, \mathrm{D}_\lambda\right), \quad \lambda_i \sim p(\lambda_i), \quad \tau \sim p(\tau), \quad \sigma \sim p(\sigma), \tag{3}$$

where D_λ is the diagonal matrix that contains the scales λ_i as diagonal elements and Ω is a correlation matrix. We refer to priors of this form (3) as dependency-aware shrinkage priors (DASP). The standard shrinkage prior model (2) is recovered by setting $\Omega = \mathrm{I}$.

2.2.1. Related priors

Priors for regression coefficients that capture dependence structures typically do so through the incorporation of a covariance matrix. In what follows, we discuss two such priors: Zellner's g prior and Structured Shrinkage Priors (SSPs), emphasizing their respective advantages and identifying their limitations. We argue that the prior introduced in (3) offers a natural and robust extension that directly addresses these limitations.

Zellner's g prior (Zellner, 1986; Maruyama and George, 2011; Li and Clyde, 2018) incorporates dependence information directly from the design matrix X by specifying

$$b \mid \sigma, g \sim \mathcal{N}\left(0, \sigma^2 g (X'X)^{-1}\right), \quad \sigma \sim p(\sigma), \tag{4}$$

where $g > 0$ is either fixed or endowed with a prior distribution, and X is the design matrix composed of the covariate vectors x_i'. This formulation ensures that the prior covariance of b is proportional to the covariance of the maximum likelihood estimator (MLE) $\hat{b}$, thus preserving scale invariance with respect to the regressors. Zellner's prior is widely appreciated for its analytical tractability, particularly in Bayesian model selection. When g is fixed, it yields closed-form

expressions for the marginal likelihood and posterior, thereby eliminating the need for sampling-based approximations (Robert and Casella, 2004; Maruyama and George, 2011; Li and Clyde, 2018).

Within this framework, g functions as a global shrinkage parameter, applying uniform regularization across all components of b. Although $(X'X)^{-1}$ provides some localized scaling, the g prior lacks adaptivity, as it does not include coefficient-specific local scales λ_i (Liang et al., 2008; Gelman et al., 2013). As a result, fixed or poorly chosen values of g can lead to over-shrinkage of relevant variables or under-penalization of noise. While various priors for g have been proposed to mitigate this (Liang et al., 2008; Maruyama and George, 2011; Li and Clyde, 2018), the absence of local scales ultimately excludes the g prior from the class of adaptive shrinkage priors. In high dimensional settings, the g prior is further limited by its reliance on $(X'X)^{-1}$, making it sensitive to collinearity and rank deficiency. By contrast, our formulation (3) is capable of mimicking the g prior through appropriate specification of Ω, as shown in Section 2.3. Moreover, the inclusion of local scale parameters λ_i directly addresses the issue of adaptivity (Liang et al., 2008), offering a flexible and robust alternative in both low and high dimensional settings.

In the broader context of shrinkage priors, a structurally related idea appears in the work of George and McCulloch (1993, 1995), who propose the spike-and-slab prior with a non-diagonal precision matrix to encode dependencies among coefficients. In their formulation, the prior on b takes the form $\mathcal{N}(0, \tau^2\Gamma\Omega^{-1}\Gamma)$, where Γ is a diagonal matrix of binary inclusion indicators $\gamma_i \in \{0, 1\}$ for $i = 1, \ldots, p$. Their work suggests using $\Omega = (X'X)^{-1}_{\gamma}$, where X_γ is the submatrix of X corresponding to the active variables, and also explores special cases such as $\Omega = I$ and $\Omega = (X'X)^{-1}$, independent of γ. However, in practice, their implementations focus on mixtures of normals with diagonal Ω, retaining conditional independence among coefficients which gives rise to the typical spike-and-slab prior. Our formulation differs in several key respects: we consider continuous local scales λ_i instead of binary indicators, yielding a fully continuous global-local prior that avoids discrete model search (Hahn and Carvalho, 2015; Tadesse and Vannucci, 2021), and we construct Ω from the full design matrix or a structured estimate thereof, allowing scalable and adaptive shrinkage in high dimensional settings. We show the latter in Section 2.3.

Turning to continuous global-local formulations, Griffin and Hoff (2024) propose the Structured Shrinkage Priors (SSPs) framework, which introduces dependence structures via a product representation of normal scale mixtures. Specifically, if $z \sim \mathcal{N}(0, \Phi)$ and λ is a vector of stochastic scales independent of z, then the elementwise product $b = \lambda \circ z$ defines a scale mixture of normals (West, 1987), with prior covariance:

$$\Sigma_b = \mathbb{E}(\lambda\lambda') \circ \Phi. \tag{5}$$

In standard settings, $\Phi = I$ yields unstructured shrinkage priors. However, when $\Phi \neq I$, identifiability issues arise: the individual scales λ_i are not separately identifiable from the diagonal entries of Φ. To address this, Griffin and Hoff (2024)

impose the constraint $\mathbb{E}(\lambda_i^2) = 1, i = 1, \ldots, p$. While this restores identifiability, it significantly restricts the flexibility of the prior by limiting the range of hyperparameter choices available for the distribution of λ_i.

For instance, under this constraint, the Normal-Gamma prior (Brown and Griffin, 2010), with $\lambda^2 \sim \text{Gamma}(\alpha, \beta)$, is only admissible when $\alpha = \beta$. Similarly, heavy-tailed priors such as the Beta Prime, which are often used to induce marginal heavy tails in b, become infeasible unless their parameter values satisfy the constraint (Pas et al., 2016; Ghosh and Chakrabarti, 2017; Bai and Ghosh, 2019; Zhang et al., 2020; Aguilar and Bürkner, 2023). This constraint also prevents straightforward generalizations of well-established priors like the Horseshoe (Carvalho et al., 2010), since it assumes $\lambda_i \sim \mathcal{C}^+(0, 1)$, implying that $\mathbb{E}(\lambda_i^2)$ does not exist.

Finally, we observe that the matrix Φ in the formulation of Structured Shrinkage Priors (SSPs) remains relatively underexplored, even though it plays a critical role in determining the properties of the model. As Griffin and Hoff (2024) illustrate, altering Φ within traditional shrinkage priors can lead to a notable increase in computational cost, emphasizing the importance of a more systematic study of its effects. In addition, the task of specifying Φ is typically left to the user, either through direct selection or by placing a prior distribution on it; both of which present open opportunities for further investigation. In contrast, our dependency-aware shrinkage prior avoids imposing any constraint on the second moment of the local scales λ_i. This flexibility allows it to encompass a broad class of shrinkage priors, including those with undefined or heavy-tailed moments, as long as Ω is appropriately specified. In the following sections, we analyze some properties of our proposed prior and examine the practical consequences of incorporating Ω.

2.2.2. Conditional means

First, we discuss how the conditional posterior mean is affected by the presence of Ω. The conditional posterior distribution of b under our dependency-aware shrinkage prior (3) is multivariate Gaussian. Its mean and covariance matrix are given by

$$\mathbb{E}(b \mid y, \lambda, \tau, \sigma, \Omega) = Q_\Omega^{-1} X' y, \;\; \text{Cov}(b \mid y, \lambda, \tau, \sigma, \Omega) = \sigma^2 Q_\Omega^{-1}, \tag{6}$$

where $Q_\Omega \coloneqq X'X + 1/\tau^2 D_\lambda^{-1} \Omega^{-1} D_\lambda^{-1}$. We explicitly condition on Ω to highlight its role in shaping both the posterior mean and covariance structure. Assuming that X is of full rank, the posterior mean can be alternatively expressed in terms of the MLE $\hat{b}$ as

$$\mathbb{E}(b \mid y, \lambda, \tau, \sigma, \Omega) = \tau^2 D_\lambda \Omega D_\lambda \left(\tau^2 D_\lambda \Omega D_\lambda + (X'X)^{-1}\right)^{-1} \hat{b}. \tag{7}$$

This expression makes explicit the nature of shrinkage applied to $\hat{b}$ through the interaction of the global scale τ, the local scales λ, and the structure-inducing

matrix Ω. To assess the specific influence of Ω, we consider the difference between the posterior means obtained under Ω and the identity matrix I:

$$\mathbb{E}(b \mid y, \lambda, \tau, \sigma, \Omega) - \mathbb{E}(b \mid y, \lambda, \tau, \sigma, I) = \left(Q_\Omega^{-1} - Q_I^{-1}\right) X'y. \tag{8}$$

This formulation allows us to isolate and quantify the effect of introducing Ω on the regularization applied to the MLE. In particular, the term $Q_\Omega^{-1} - Q_I^{-1}$ captures the deviation from standard isotropic shrinkage, revealing how structured dependence alters the posterior behavior.

Let A be an invertible matrix of dimension p. Denote by $\|A\|_2$ the spectral norm of A, i.e. its largest singular value. We have derived the following two-sided bound for the spectral norm difference between Q_Ω^{-1} and Q_I^{-1} (see Appendix 5 for details).

$$\frac{\|\Omega^{-1} - I\|_2}{\lambda_1^2 \left(\nu_1 + \frac{1}{\lambda_p^2 \omega_p}\right)\left(\nu_1 + \frac{1}{\lambda_p^2}\right)} \leq \|Q_\Omega^{-1} - Q_I^{-1}\|_2 \leq \frac{\|\Omega^{-1} - I\|_2}{\lambda_p^2 \left(\nu_p + \frac{1}{\lambda_1^2 \omega_1}\right)\left(\nu_p + \frac{1}{\lambda_1^2}\right)}. \tag{9}$$

Here, λ_1, ω_1, and ν_1 denote the largest local scale, the largest eigenvalue of Ω, and the largest eigenvalue of $X'X$, respectively, while λ_p, ω_p, and ν_p denote the corresponding smallest values. We assume $\tau^2 = 1$ for notational simplicity. These inequalities quantify how structural deviations in the prior covariance, the spread of local shrinkage parameters, and the conditioning of the design matrix influence the posterior covariance matrix. In particular, the difference between the structured and independent posterior covariances is amplified when shrinkage is globally strong (i.e., small λ_1 and hence small λ_p), the design is poorly conditioned (i.e., small ν_p), or the prior structure deviates substantially from independence (i.e. large $\|\Omega^{-1} - I\|_2$). Conversely, the upper bound shows that the difference can be negligible when shrinkage is weak (i.e., large λ_p and hence large λ_1), the design is well-conditioned (i.e. large ν_p), and the structural deviation is modest (i.e. small $\|\Omega^{-1} - I\|_2$). In this regime, the dependence structure plays a limited role, and the posterior mean is largely data driven.

In particular, dependency-aware shrinkage acts in a rotated and rescaled coordinate system determined by the dependence structure Ω. To see this, let $\Omega = U\Lambda U'$, denote the eigendecomposition of Ω, and define the transformed coefficients $\tilde{b} = U'D_\lambda^{-1}b$. Conditional on τ, λ, it follows immediately that $\tilde{b} \mid \tau, \lambda \sim \mathcal{N}(0, \tau^2\Lambda)$. When Ω encodes interpretable structure, such as sparsity or neighborhood dependencies, our prior reweights the contribution of each component in the parameter space.

2.2.3. Divergences

We analyze how the inclusion of Ω alters our model by studying the divergence between probability measures, while keeping all other components fixed. In particular, we focus on the Kullback–Leibler (KL) divergence, which quantifies the discrepancy between two probability distributions. Divergence-based

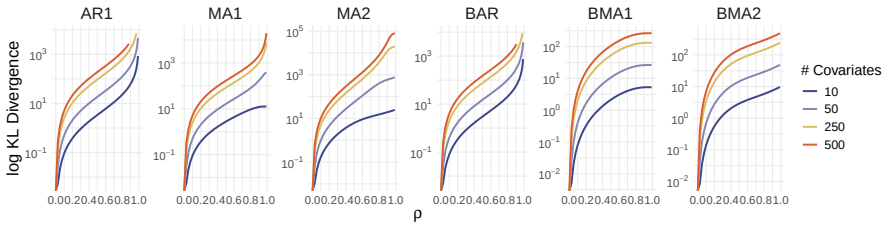


FIG 1. *KL divergence as a function of the correlation parameter ρ for various structures of Ω. The corresponding algebraic forms are provided in Appendix 5.*

metrics allow us to account for the entire distribution, rather than relying solely on pointwise comparisons.

Let $\mathcal{P} \sim \mathcal{N}(\mu_{\mathcal{P}}, \Sigma_{\mathcal{P}})$ and $\mathcal{Q} \sim \mathcal{N}(\mu_{\mathcal{Q}}, \Sigma_{\mathcal{Q}})$ be two p dimensional multivariate normal distributions. Their KL divergence has the following closed-form expression:

$$\mathrm{KL}\left(\mathcal{P} \,||\, \mathcal{Q}\right) = \frac{1}{2}\left[(\mu_{\mathcal{Q}} - \mu_{\mathcal{P}})'\Sigma_{\mathcal{Q}}^{-1}(\mu_{\mathcal{Q}} - \mu_{\mathcal{P}}) + \mathrm{tr}\left(\Sigma_{\mathcal{Q}}^{-1}\Sigma_{\mathcal{P}}\right) - \ln\frac{|\Sigma_{\mathcal{P}}|}{|\Sigma_{\mathcal{Q}}|} - p\right]. \tag{10}$$

Now consider $\mathcal{P}$ as the standard shrinkage prior for b with $\Omega = I$ (see (2)), and let $\mathcal{Q}$ be a prior of the form (3). The Kullback–Leibler (KL) divergence between the conditional prior distributions for $b \mid \sigma, \tau, \lambda$ depends explicitly on Ω and is given by

$$\mathrm{KL}\left(\mathcal{P} \,||\, \mathcal{Q}\right) = 1/2\left(\mathrm{tr}\left(\Omega^{-1}\right) + \ln|\Omega| - p\right). \tag{11}$$

This expression emphasizes how Ω influences the relative entropy between the two priors. The trace term $\mathrm{tr}(\Omega^{-1})$ indicates that nearly ill-conditioned (or nearly singular) matrices lead to a large KL divergence. The term $\ln|\Omega|$ captures the overall volume scaling of the covariance structure imposed by Ω.

Figure 1 illustrates how the KL divergence varies across different specifications of Ω. In the provided examples, each correlation matrix Ω is parameterized by a single hyperparameter $\rho \in (0, 1)$, which fully determines its structure. We consider commonly used correlation patterns, including the autoregressive model of order 1 (AR1), moving average models of orders 1 and 2 (MA1 and MA2), as well as their blocked counterparts: BAR1, BMA1, and BMA2. These same structures are also employed in our experiments in Section 3. We provide their algebraic definitions in Appendix 5.

The results in Figure 1 show that the KL divergence increases monotonically with the number of covariates, and the discrepancy between the conditional distributions becomes more pronounced as ρ increases—especially with a sharp rise around $\rho \approx 0.9$. This indicates that differences between prior specifications are most substantial when parameters are strongly correlated. Consequently, we conjecture that posterior inference under the DASP will diverge most noticeably from standard shrinkage priors in high-correlation settings.

2.2.4. Joint concentration and joint tail behavior

Most shrinkage priors are discussed through one-dimensional marginals $p(b_i)$, because conditional independence makes the joint distribution difficult to interpret. In DASP, dependence is introduced explicitly, so the joint marginal prior

$$p(b) = \int p(b \mid \sigma, \tau, \lambda, \Omega)\, p(\lambda)\, p(\tau)\, p(\sigma)\, d\lambda\, d\tau\, d\sigma \tag{12}$$

carries additional information that is not visible from the marginals alone. In particular, matching the same one-dimensional marginals does not determine how the prior behaves near the origin jointly, or how much mass it assigns to regions where several coefficients are simultaneously large.

To make this distinction explicit, we study a common-scale special case where $b = \lambda z$ with $z \sim \mathcal{N}(0, \Omega)$ and $\lambda \sim \mathrm{C}^+(0, 1)$ independent of z. Using the same λ for every coefficient isolates the effect of Ω on the joint shape of the marginal prior. In fact, each component b_i has the same one-dimensional marginal density $p(\cdot)$ for any correlation matrix Ω, so differences in prior behavior are due to the inclusion of the dependence structure. We compare the joint density $p_\Omega(b)$ to the product construction

$$p_I(b) = \prod_{i=1}^{p} p(b_i), \tag{13}$$

which has the same one-dimensional marginals but imposes independence.

Proposition 1 (Joint spike and joint tail mass under a common-scale special case)**.** *Let $p \geq 2$ and let Ω be a positive definite correlation matrix in $\mathbb{R}^{p\times p}$. Let $\lambda \sim \mathrm{C}^+(0, 1)$ be independent of $z \sim \mathcal{N}(0, \Omega)$ and define $b = \lambda z$. Write $p_\Omega(b)$ for the joint density of b, and let $p(\cdot)$ denote the common one-dimensional marginal density of each b_i. Define $p_I(b) = \prod_{i=1}^{p} p(b_i)$. Then the following hold.*

(i) As $b \to 0$,

$$p_\Omega(b) \asymp \left(b'\Omega^{-1}b\right)^{-(p-1)/2}, \tag{14}$$

and in particular $p_\Omega(b)/p_I(b) \to \infty$.

(ii) Let $A = \{x \in \mathbb{R}^p : \min_{1\leq i\leq p} |x_i| > 1\}$. Then

$$P(b \in tA) \sim \frac{2}{\pi} \frac{\mathbb{E}\left[\min_{1\leq i\leq p} |z_i|\right]}{t} \qquad \text{as } t \to \infty, \tag{15}$$

while under $p(\cdot)$ one has $P_I(b \in tA) \asymp t^{-p}$. Consequently, $P(b \in tA)/P_I(b \in tA) \to \infty$.

Proposition 1 shows two joint effects that are not determined by the marginals. Part (i) shows that introducing dependence through Ω can strengthen concentration near $b = 0$ under the joint prior, even though the one-dimensional marginals are unchanged. Part (ii) shows that this is not obtained by lightening tails. In the same setting, the dependent construction can assign more mass to joint tail regions than the product construction with the same marginals, so simultaneous large coefficients are not ruled out.

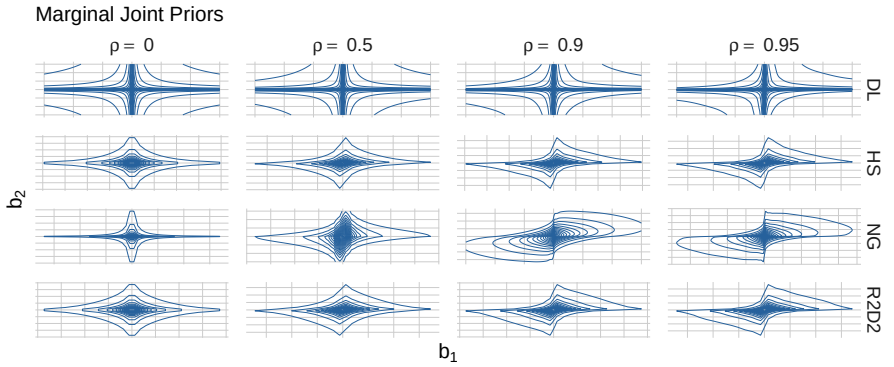


FIG 2. *Monte Carlo approximations of the bivariate joint marginal prior distribution* $p(b_1, b_2)$ *under different shrinkage priors: Dirichlet-Laplace (DL), Horseshoe (HS), Normal-Gamma (NG), and R2D2, and for various correlation levels* ρ*. We use default hyperparameters for each prior, with* $\sigma = 1$*, and* $\Omega = (1-\rho)I + \rho JJ'$*, where* J *is a* 2*-dimensional vector of ones.*

2.2.5. Contour plots of bivariate marginals

The previous subsection shows that dependence can change the joint behavior of the marginal prior in ways that are not visible from $p(b_i)$. We now illustrate the resulting geometry through contour plots of the bivariate marginal $p(b_1, b_2)$ under several base priors and correlation levels. Since $p(b_1, b_2)$ is not available in closed form, we approximate it by Monte Carlo (Robert and Casella, 2004).

Figure 2 illustrates how introducing a dependency structure through Ω influences the joint prior distribution $p(b)$. In this analysis, we fix $\sigma = 1, p = 2$ and set $\Omega = (1-\rho)I + \rho JJ'$, where J is a 2-dimensional vector of ones, varying the value of ρ to control the degree of correlation. We apply this procedure to several widely used shrinkage priors: Dirichlet–Laplace (DL), Horseshoe (HS), Normal–Gamma (NG), and R2D2 (Bhattacharya et al., 2015; Carvalho et al., 2010; Brown and Griffin, 2010; Zhang et al., 2020), using default hyperparameter values as recommended in the literature. See Appendix 5 for details on the priors and their default hyperparameters.

The matrix Ω enters through the Gaussian distribution $p(b \mid \lambda, \tau)$ and determines the orientation of $b = (b_1, b_2)'$. After marginalizing over the scales, the joint marginal $p(b)$ need not preserve this dependence structure. In particular, $p(b)$ is typically non-Gaussian even when $p(b \mid \lambda, \tau)$ is Gaussian, so conditional correlation does not translate directly into the same marginal correlation. Figure 2 shows how introducing Ω affects the bivariate marginal shape across different values of ρ. When $\rho = 0$, the contours concentrate near the coordinate axes, which reflects prior mass on sparse configurations where one coefficient is near zero while the other can be nonzero. As ρ increases, prior mass shifts toward the main diagonal, so the bivariate marginal becomes more diagonal-aligned and less concentrated near the coordinate axes. This indicates weaker joint preference for configurations in which one coefficient is close to zero

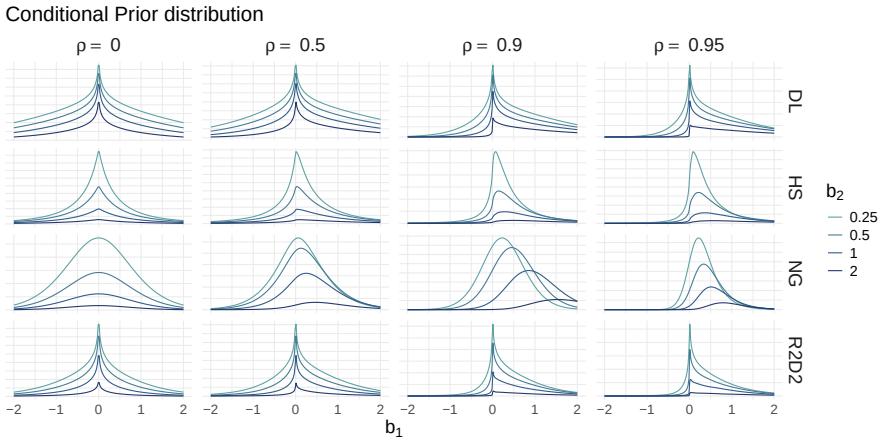


FIG 3. *Monte Carlo approximations of the unnormalized conditional prior distribution* $p(b_1 \mid b_2)$ *for varying values of* b_2 *and* ρ*. As both* b_2 *and* ρ *increase, the distribution shifts toward larger values of* b_1*, indicating reduced shrinkage.*

while the other is not, for fixed hyperparameters. At high correlation, the prior increasingly favors $b_1 \approx b_2$, which resembles the pooling behavior induced by ℓ_2-type regularization for highly collinear, similarly scaled predictors (Hoerl and Kennard, 1970).

2.2.6. Conditional distributions

Figure 3 displays Monte Carlo approximations of the conditional distributions $p(b_1 \mid b_2)$, using the same covariance structure Ω as in previous figures, while varying ρ and b_2. When $\rho = 0$, increasing b_2 leads to a marginal distribution for b_1 with reduced mass near zero. As both ρ and b_2 increase, the conditional distribution shifts markedly to the right, with b_2 primarily determining the magnitude of this shift. The pronounced spikes in the plot arise because we are visualizing $p(b_1 \mid b_2)$ rather than the fully conditional $p(b_1 \mid b_2, \lambda, \tau)$, which are Gaussian. Consistent with the contour plots in Figure 2, the marginal distribution becomes smoother as ρ increases, suggesting that less shrinkage is induced on b_1 in the presence of stronger correlation.

2.2.7. Shrinkage properties

Shrinkage properties can be studied through shrinkage factors, which quantify how strongly each coefficient b_i is pulled toward zero (Carvalho et al., 2010; Piironen and Vehtari, 2017; Aguilar and Bürkner, 2023). The well-known normal means problem arises in Model (1) when the design matrix is the identity, i.e., $X = I$ (Castillo and Vaart, 2012). Under a standard shrinkage prior, an

application of Fubini's theorem yields $\mathbb{E}(b_i \mid y_i, \lambda_i) = (1 - \kappa_i)y_i$, where the quantity

$$\kappa_i := \frac{1}{1 + \lambda_i^2}, \tag{16}$$

is referred to as the shrinkage factor (Carvalho et al., 2010; Piironen and Vehtari, 2017; Bai and Ghosh, 2019; Aguilar and Bürkner, 2023). It represents the proportion of the posterior mean of b_i that is attributed to shrinkage toward zero after observing the data y. Since $0 \leq \kappa_i \leq 1$, the law of total expectation implies $|\mathbb{E}(b_i|y_i)| = |(1 - \mathbb{E}(\kappa_i|y_i))y_i| \leq |y_i|$. In this context, y_i serves as MLE of b_i, so κ_i can be interpreted as the degree to which the MLE is shrunk toward zero. When b_i corresponds to signal (i.e., a large effect), $\kappa_i \approx 0$, whereas for noise, $\kappa_i \approx 1$.

Shrinkage factors offer a unified framework for comparing the behavior of different shrinkage priors across a range of hyperparameter settings (Tadesse and Vannucci, 2021). A desirable property of such priors is their ability to effectively distinguish between noise and signal components. This can be studied by examining the prior distribution of the shrinkage factor κ_i, which reveals how aggressively the prior shrinks coefficients under varying hyperparameter configurations. For instance, the Horseshoe prior shows a bimodal behavior in the implied distribution for $\kappa_i \sim \text{Beta}(1/2, 1/2)$, with mass concentrated near 0 and 1 (Carvalho et al., 2010; Armagan et al., 2011; Pas et al., 2017). This enables the prior to strongly shrink noise (large κ_i) while retaining signals (small κ_i).

The notion of shrinkage factors from Equation (16) can be extended to settings with nontrivial dependency structures ($\Omega \neq I$) by examining the conditional posterior mean of $b \mid \sigma, \tau, \lambda$ given in Equation (6). This allows us to assess how the posterior mean deviates from the MLE under more general dependence structures. Specifically, we define the matrix-valued shrinkage factor as:

$$\kappa_\Omega = I - \tau^2 D_\lambda \Omega D_\lambda (\tau^2 D_\lambda \Omega D_\lambda + (X'X)^{-1})^{-1}. \tag{17}$$

This definition becomes cumbersome when approaching shrinkage factors from the perspective of MLE shrinkage. Unlike the classical scalar shrinkage factors, κ_Ω is matrix-valued and generally non-diagonal, reflecting the induced dependence across coefficients due to Ω. To illustrate this, consider the special case with $\tau = \sigma = 1$, $\Omega = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$, and $X = I$. In this case, the conditional posterior mean becomes:

$$\mathbb{E}(b \mid y, \lambda) = \frac{1}{1 + \lambda_1^2 + \lambda_2^2 + \lambda_1^2\lambda_2^2(1 - \rho^2)} \begin{pmatrix} \lambda_1^2(1 + \lambda_2^2(1 - \rho^2))\hat{b}_1 + \rho\lambda_1\lambda_2\hat{b}_2 \\ \lambda_2^2(1 + \lambda_1^2(1 - \rho^2))\hat{b}_2 + \rho\lambda_1\lambda_2\hat{b}_1 \end{pmatrix}, \tag{18}$$

where $\hat{b}_1$ and $\hat{b}_2$ are the MLEs of b_1, b_2 respectively. When $\rho = 0$, κ_Ω reduces to a diagonal matrix, and we recover the classical shrinkage factors $\kappa_i = \frac{1}{1+\lambda_i^2}$ by taking the diagonal. The correlation parameter ρ controls the degree of interaction between the local scales λ_1 and λ_2. As ρ increases, the posterior means of b_1 and b_2 become more entangled, and shrinkage toward zero weakens unless

the MLEs are already near zero. This also reflects a global pooling effect, where coefficients are increasingly pulled toward each other. Hence posterior estimates are pulled not only toward zero, but also toward one another.

Importantly, the introduction of correlation through Ω does not eliminate shrinkage toward zero. Instead, it makes shrinkage less aggressive in the direction of differences between MLEs, especially when both estimates are large in magnitude. The correlation parameter ρ controls a trade-off between promoting sparsity (via local shrinkage through λ_i) and encouraging similarity across coefficients (via pooling). Full shrinkage remains possible due to the continued flexibility of the local scales λ_i.

While shrinkage factors offer a local, coefficient-wise view of regularization, a more global perspective is provided by the effective number of nonzero coefficients, m_{eff}, introduced by Piironen and Vehtari (2017) and defined as:

$$m_{\text{eff}} = \sum_{i=1}^{p} (1 - \kappa_i). \tag{19}$$

This quantity serves as a measure of the model's effective dimensionality, satisfying $m_{\text{eff}} \leq p$, and reflects how the combination of local and global shrinkage affects model complexity. When prior information or domain expertise suggests a plausible number of active predictors, visualizing the prior distribution of m_{eff} under different hyperparameter settings offers practical guidance for prior elicitation (Piironen and Vehtari, 2017; Aguilar and Bürkner, 2023). The user can rely on m_{eff} to match the expected model complexity to domain knowledge without needing to examine each λ_i individually. This approach is especially informative in high-dimensional regimes, where balancing flexibility and parsimony is critical. We extend this definition to our dependency-aware shrinkage prior by using the matrix valued shrinkage factor κ_Ω, leading to

$$m_{\text{eff}} = \text{tr}(I - \kappa_\Omega). \tag{20}$$

This extension preserves the original interpretation: the trace still quantifies the influence of the data on the posterior mean, but now incorporates the dependence structure among coefficients induced by Ω. In other words, m_{eff} continues to represent an effective model size, but under a prior where shrinkage is applied not just individually, but collectively. This generalization is particularly meaningful because it reflects how correlation alters the balance between sparsity (zero shrinkage) and similarity (pooling). Notably, we recover the classical definition when $\Omega = I$.

The posterior distribution of m_{eff} serves as a useful diagnostic to monitor how the model adapts to the data. Comparing posterior expectations of m_{eff} across different prior choices (e.g., independent vs. correlated shrinkage) helps quantify how strongly the priors influence model complexity in practice, and to what extent dependencies in Ω are effectively "used" in posterior inference. We illustrate the prior distribution of m_{eff} for various shrinkage priors in Figure 4. For these simulations, we fixed $p = 100$, used $X = I$, and employed standard

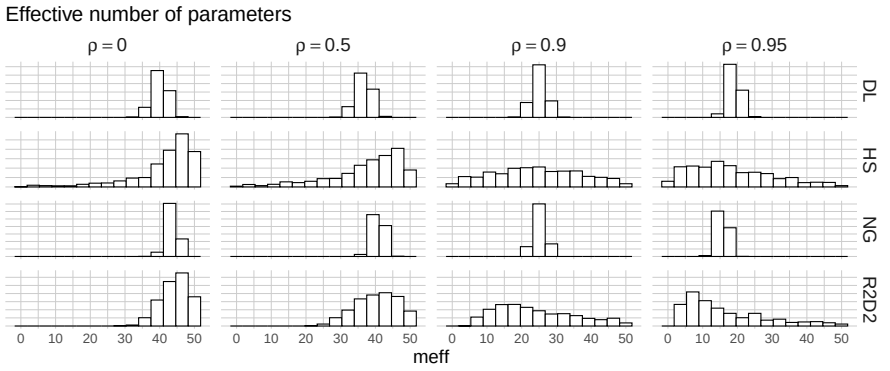


FIG 4. *Implied priors of the effective number of parameters (Equation (20)) for different shrinkage priors (rows) and correlations (columns). We have used the default hyperparameters for the different shrinkage priors, set* $p = 100$ *covariates,* $\sigma = 1$ *and* $\Omega = (1 - \rho)I + \rho JJ'$*, where* J *is a* p*-dimensional vector of ones.*

hyperparameters for each prior. The correlation structure was specified as $\Omega = (1 - \rho)I + \rho JJ'$, where J is a p-dimensional vector of ones, and we varied ρ to assess the impact of correlation. As ρ increases, the effective model size m_{eff} grows, indicating less global shrinkage. This is due to the increased dependence among the coefficients, which results in increased pooling. However, this effect can be moderated by adjusting the hyperparameters responsible for shrinkage, such as increasing concentration near zero, which would counterbalance the growth in effective model size induced by the correlation.

2.3. Specification of dependence structures

A significant challenge in the use of model (3) is the specification of the matrix Ω. We discuss our preferred approach to specifying Ω. Potential alternative methods are detailed in the Discussion section.

It is possible to extract correlational information about the regression coefficients b directly from the design matrix X. Since X encodes information about how the predictors relate to each other, we can model the dependencies between the regression coefficients b based on the structure of X. A well-known example of this approach is Zellner's g prior (see Section 2.2.1), which achieves this by setting the prior covariance matrix of the coefficients proportional to the covariance of the MLE, i.e., $\text{Cov}(b \mid \sigma, g) = g\,\sigma^2(X'X)^{-1}$ where $g > 0$. An alternative interpretation of Zellner's prior is that it uses a covariance structure for b that is proportional to the true covariance matrix of the predictors, Σ_X. This is because when n is sufficiently large, the unbiased estimator of the covariance of X is approximately $X'X/n$, which converges to Σ_X as $n \to \infty$ and p is fixed.

Inspired by Zellner's prior, we propose a generalization of shrinkage priors that incorporates the structure of the design matrix X, aligning with the prior structure of our proposed model in Equation (3). Specifically, we derive the

dependence structure Ω directly from X, using its covariance matrix Σ_X to inform the relationships between the regression coefficients. Let $\text{Cov}(X) = \Sigma_X$ denote the true covariance matrix of X. To define Ω, we first compute the inverse of Σ_X, i.e., the precision matrix of X, which we denote by $\Theta_X = \Sigma_X^{-1}$. We then standardize Θ_X to obtain a correlation matrix, which we use as Ω. Explicitly, we write:

$$\Omega = \text{Cor}(\Theta_X) = \text{Cor}\left(\Sigma_X^{-1}\right), \tag{21}$$

where the Cor operator is the normalized precision matrix, whose off-diagonal entries equal the negatives of the estimated partial correlations. Therefore, Ω encodes the correlation structure between the regression coefficients, reflecting the dependencies among them based on the design matrix X. This leads to the following prior covariance for b:

$$\text{Cov}(b \mid \lambda, \tau, \sigma) = \sigma^2 \tau^2 D_\lambda \, \text{Cor}\left(\Sigma_X^{-1}\right) D_\lambda. \tag{22}$$

This choice of Ω has several key interpretations:

1. **Relationship to the partial precisions**. Let Ψ_X denote the partial precision matrix of X (Lam, 2020). Under our formulation, we have

   $$\Omega_{ij} = -\Psi_{ij}, \quad i \neq j.$$

   which implies that Ω_{ij} encodes the conditional correlation between b_i and b_j after adjusting for all other variables. Since b follows a conditional multivariate Gaussian distribution, whenever $\Omega_{ij} = 0$, the coefficients b_i and b_j are conditionally uncorrelated given all other variables (Giraud, 2014). This structure allows us to model the dependencies between the regression coefficients directly from the design matrix. The relationship to partial precision matrices emphasizes how Ω reflects conditional dependencies between coefficients.
2. **Prior-Likelihood agreement**. The conditional posterior mean of $b \mid \sigma, \tau, \lambda$ (see Equation (7)) suggests that aligning the prior correlation structure of Ω with the correlation structure implied by $(X'X)^{-1}$ enhances the coherence between the prior and the likelihood (Zellner, 1996). Specifically, both the prior and the likelihood would then imply the same posterior correlation structure. In principle, this alignment should reduce prior-likelihood conflict and facilitate inference.
3. **Imaginary observations**. Another interpretation of this prior structure is in terms of imaginary or pseudo-observations. Specifically, the model can be viewed as augmenting the data with additional pseudo observations that share the same design matrix X, but with response values set to zero (Zellner, 1986). These pseudo-observations are scaled by the coefficients λ_i and τ, which determine their relative weight.

In practice, the true covariance matrix Σ_X is typically unknown. Instead, it must be estimated from the observed design matrix X, which can be particularly challenging when $p > n$. We discuss methods for estimating Σ_X in Section 2.4.

By using a suitable estimator for Σ_X, we can proceed with standard Bayesian inference. Importantly, since the entire regression model is conditioned on X, allowing the prior to depend on it poses no issues in terms of consistency or model specification.

2.4. Estimating the dependence structure Ω

In many applications, the predictor covariance Σ_X is unknown and must be approximated from the design matrix X (typically after centering and scaling the columns). Our construction only requires a positive definite correlation matrix Ω, and in principle Ω may be supplied externally (for example from known groupings or scientific structure). When Ω is learned from X, the main practical issue is numerical stability under collinearity and when p is not small relative to n.

Let

$$S = \frac{1}{n-1} X^\top X$$

denote the sample covariance estimator of Σ_X. When $p < n$ and S is well-conditioned, one may form a dependence matrix from the inverse covariance by setting

$$\Omega = \mathrm{Cor}\left(S^{-1}\right), \tag{23}$$

which uses the partial-correlation structure implied by X. However, under strong multicollinearity, S can be ill-conditioned even when it is invertible, making (23) unstable. For this reason, we treat (23) as a conceptual baseline and recommend regularized estimators in practice.

In particular, when p is large relative to n, regularization is essential. Under proportional asymptotics with $p/n \to c \in (0, \infty)$, the sample covariance S is not a consistent estimator of Σ_X in general, and its eigenvalues do not concentrate even when $\Sigma_X = I$ (Marčenko and Pastur, 1967; Ledoit and Wolf, 2004a,b; Lam, 2020). This further motivates replacing S by a well-conditioned estimator.

Among the many available approaches, we focus on estimators that require minimal user input and return a positive definite matrix without imposing strong structural assumptions on Σ_X. A standard choice is linear shrinkage covariance estimation (Ledoit and Wolf, 2004b). The population shrinkage target within this class is

$$\Sigma^* = \varphi_1 I + \varphi_2 S = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2} S, \tag{24}$$

where $\delta^2 = \alpha^2 + \beta^2$, $\mu = \mathrm{tr}(S)/p$, and the coefficients are chosen to minimize the expected Frobenius loss over the class $\{\varphi_1 I + \varphi_2 S\}$ (Ledoit and Wolf, 2004b). Since Σ_X is unknown, one uses consistent estimators for μ, α^2, and β^2 to obtain an empirical shrinkage estimator S^* (see Appendix 5). The estimator S^* is positive definite, computationally simple, and consistent under high-dimensional asymptotics where p and n grow at a proportional rate (Silverstein, 1995; Ledoit and Wolf, 2004b). It preserves the eigenvectors of S while shrinking eigenvalues toward a multiple of the identity matrix (Ledoit and Wolf, 2004a,b).

We then define

$$\Omega = \mathrm{Cor}\big((S^*)^{-1}\big)\,, \tag{25}$$

and use this Ω in the dependency aware construction.

Finally, we emphasize that (25) is a baseline default. A large literature studies regularized estimation of high-dimensional covariance and precision matrices, including Bayesian procedures with posterior contraction guarantees under structural assumptions such as sparsity or factor structure (Pati et al., 2014; Banerjee and Ghosal, 2014; Li et al., 2019; Banerjee et al., 2021; Lee et al., 2022; Lee and Lee, 2021; Sagar et al., 2024). Any procedure that returns a positive definite estimate of either Σ_X or Σ_X^{-1} can be used to construct Ω. Concretely, one may set $\Omega = \mathrm{Cor}(\widehat{\Sigma}^{-1})$ for a regularized covariance estimator $\widehat{\Sigma}$, or equivalently $\Omega = \mathrm{Cor}(\widehat{\Theta})$ for a regularized precision estimator $\widehat{\Theta} \approx \Sigma_X^{-1}$. Furthermore, one can combine several estimators to guard against misspecification of the dependence structure. Namely if we have a family of estimators$\{I_p, \Omega^{(1)}, \ldots, \Omega^{(K)}\}$, we can pool them via $\sum_i \eta_i \Omega^{(i)}$, where $\Omega^{(0)} = I_p$ and the weights $\eta_i \geq 0$ and $\sum_i \eta_i = 1$ are to be determined. This can be done by Bayesian procedures or empirical Bayes. In general, the appropriate choice depends on which structural assumptions about Σ_X are plausible for the application, on computational constraints, and, most importantly, on placing prior variance in directions where signal is plausible. In our experiments we use Ledoit–Wolf shrinkage to avoid additional tuning parameters and to keep the default construction broadly applicable.

2.5. Workflow of the Method

We briefly summarize the practical steps for applying a dependency-aware shrinkage prior. The procedure begins by selecting a standard shrinkage prior. If the covariance matrix Σ_X is known or specified by the user, the dependence matrix Ω is computed directly via Equation (21). When Σ_X is unknown and $p < n$, Ω is constructed using the sample covariance matrix as per Equation (23). In high dimensional settings ($p > n$) or when X is ill-conditioned, we recommend estimating Σ_X using the Ledoit–Wolf shrinkage estimator S^*, Equation (24), and computing Ω via Equation (25). This workflow integrates dependence information into the prior while maintaining numerical stability. We adopt this procedure in the experiments presented in Section 3.

3. Experiments

3.1. Simulations

We conducted simulation studies to evaluate the impact of incorporating dependency structures in shrinkage priors on model performance. We examined whether encoding dependence via correlation matrices, as discussed in Section 2, influences parameter recovery and predictive performance. Our simulation

conditions reflect realistic scenarios encountered in practice. Additionally, we analyze real world datasets commonly used as benchmarks for dependence-structured models, allowing us to compare commonly observed empirical covariance structures with those assumed in our simulations.

We considered the following shrinkage priors: Beta Prime (BP), Dirichlet-Laplace (DL), Horseshoe (HS), Regularized Horseshoe (RHS), and R2D2 (D2) (Bai and Ghosh, 2019; Carvalho et al., 2010; Bhattacharya et al., 2015; Piironen and Vehtari, 2017; Zhang et al., 2020). While this list is not exhaustive, it adequately represents the main approaches explored in shrinkage prior research. Given that we are interested in the sparse scenario, we have used default hyperparameters that imply strong shrinkage as recommended in the literature. See Appendix 5 for details.

All models were implemented in Stan (Carpenter et al., 2017; Stan Development Team, 2024), which employs an adaptive Hamiltonian Monte Carlo (HMC) sampler known as the No-U-Turn Sampler (NUTS) (Neal, 2011; Brooks et al., 2011; Hoffman and Gelman, 2014) to sample draws from posterior distributions. The associated data and code are available at `https://osf.io/fuean/`.

3.1.1. Evaluation Metrics

To assess the effect of incorporating correlation structures into shrinkage priors, we compare models with and without dependence structures. Let M denote a model with a standard shrinkage prior and M_Ω the same model incorporating a correlation matrix Ω. Since the only difference between these models is the inclusion of Ω, any variation in performance can be attributed to its presence. For a given quantity of interest $\mathcal{Q}$, we define the change in performance as $\Delta\mathcal{Q} = \mathcal{Q}(M_\Omega) - \mathcal{Q}(M)$. We evaluate model performance using the following metrics: out-of-sample predictive performance, parameter recovery and coverage.

Out-of-Sample Predictive Performance: Predictive accuracy was assessed using the expected log-pointwise predictive density (ELPD) (Vehtari and Ojanen, 2012; Vehtari et al., 2016), computed as $\text{ELPD} = \sum_{i=1}^{N_{\text{new}}} \ln\left(\frac{1}{S}\sum_{s=1}^{S} p(y_{\text{new},i} \mid \theta^{(s)})\right)$, where $\theta^{(s)}$ represents the sth draw from the posterior distribution $p(\theta \mid y)$ for $s = 1, \ldots, S$. The ELPD quantifies predictive performance across a set of N_{new} observations unseen during model training. Higher ELPD values indicate better accuracy (Vehtari and Ojanen, 2012). Log-probability scores are recommended as a default Bayesian evaluation metric (Vehtari et al., 2016).

Parameter Recovery: We assess parameter recovery by computing the posterior Root Mean Squared Error (RMSE) for the regression coefficients b, defined as $\text{RMSE} = \frac{1}{p}\sum_{i=1}^{p}\sqrt{\frac{1}{S}\sum_{s=1}^{S}\left(b_i^{(s)} - b_i\right)^2}$, where $b_i^{(s)}$ is the sth posterior draw for coefficient b_i, and b_i denotes its true value. RMSE provides a global measure of estimation error, naturally capturing the bias-variance tradeoff (Robert, 2007). To gain deeper insights, we compute three variations of RMSE: 1) Overall RMSE (averaged across all coefficients). 2) RMSE for truly zero coefficients

(measuring the ability to shrink the effect of irrelevant predictors). 3) RMSE for truly nonzero coefficients (assessing how well true signals are recovered).

Coverage: We examine coverage properties by using 95% marginal credibility intervals. We report average coverage proportion, average width, sensitivity (power), specificity, and coverage of non zeros (Berger, 1985; Benjamini and Hochberg, 1995). We also show Receiver operating characteristic (ROC) curves to understand how coverage properties would change as we modify the size of the credibility interval. See Appendix 5 for results.

Convergence Diagnostics: To ensure reliable posterior inference, we report two key MCMC diagnostics: 1) The $\hat{R}$ statistic (Vehtari et al., 2021), which compares within-chain and between-chain variance with values close to 1.0 indicating convergence. 2) The Effective Sample Size (ESS), which measures the number of effectively independent samples. Higher values indicate lower autocorrelation and better mixing (Brooks et al., 2011).

3.1.2. Generative Models

We use model (1) as the generative model for our simulations, adapting it to accommodate various types of data encountered in practice. The design matrix X is drawn from a zero-mean multivariate normal distribution with covariance Σ_X, generated from one of the following: (1) AR1: autoregressive of order 1; (2) MAq : moving average processes with $q \in \{1, 2\}$; and (3) BAR1 and BMAq: blocked versions using nonoverlapping blocks of size 5 with zero off-block correlations. All structures are parameterized by $\rho \in \{0.5, 0.95\}$, which governs the correlation strength. See Appendix 5 for details. We fix $n = 100$ while varying $p \in \{50, 250\}$ to study both low and high dimensional settings. The residual variance σ^2 is calibrated to yield $R_0^2 \in \{0.2, 0.8\}$ (Gelman et al., 2013). We construct Ω using either the true Σ_X or its estimate: the sample covariance S when $p < n$, and the Ledoit–Wolf shrinkage estimator $\tilde{S}$ when $p > n$.

We generate the regression coefficients b_i under two schemes commonly used in the shrinkage prior literature (Carvalho et al., 2010; Brown and Griffin, 2010; Bhattacharya et al., 2015; Zhang et al., 2020). 1) In the *Block Coefficients* setup, the first and last five entries of b are set to $b^* \in \{3, 7\}$, with all others set to zero. 2) In the *Random Block Coefficients* setup, the same blocks are sampled from $\mathcal{N}(0, \Sigma_b)$, while the rest remain zero. We consider two versions of Σ_b: (A) diagonal with variance 9, and (B) AR(1) with $\rho_b = 0.8$ and the same variance. Sparsity is induced by setting each coefficient to zero with probability 0.75. Lastly, we sample the intercept as $\alpha \sim \mathcal{N}(0, 3)$.

3.1.3. Results

We focus on the subset of simulations using the blocked versions of the correlation matrix with the true value of Σ_X. The results presented here are representative of the full results, which can be found on OSF (`https://osf.io/fuean/`).

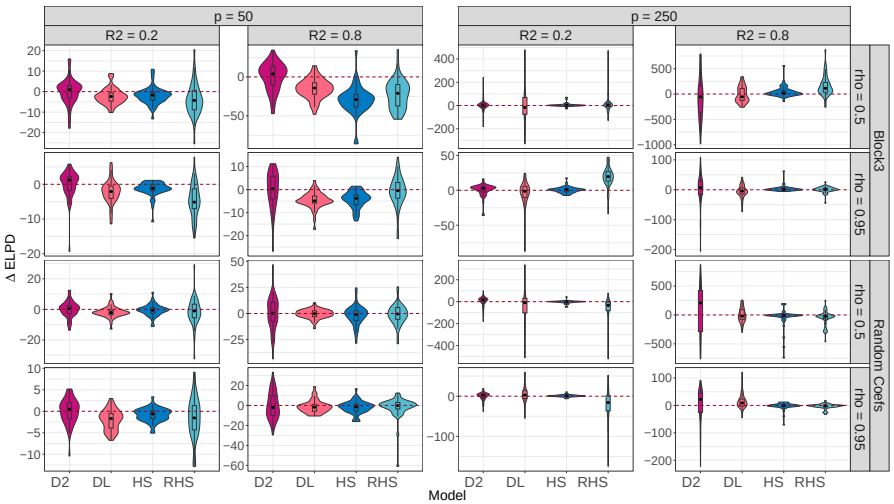


FIG 5. ***ΔELPD for each model under the BAR1 structure across simulation scenarios.*** *Model names represent the difference between the dependency-aware and standard versions, where positive values indicate improved predictive performance of the dependency-aware versions. We have omitted the BP prior due to high variability.*

We hypothesize that our dependency-aware shrinkage priors are most effective when groups (or blocks) of highly correlated predictors collectively carry strong signal, i.e., when several correlated covariates each contribute substantially to predicting the response. In such cases, encoding correlation in the prior should allow for more coherent regularization across coefficients, improving parameter recovery under structured sparsity. However, a particularly challenging scenario arises when strong signal predictors are highly correlated with irrelevant predictors. Here, the prior faces a tension between two competing goals: promoting sparsity by shrinking noise coefficients to zero, and preserving signal strength by allowing large coefficients to escape shrinkage. This trade-off can lead to suboptimal inference—either by over-shrinking the signal due to its proximity to noise, or by inflating noise variables that inherit mass from nearby signals.

Figures 5 and 6 display the results of the BAR1 and BMA1 experiments, respectively, and illustrate the impact of incorporating the dependency matrix Ω on predictive performance. Overall, the findings indicate that predictive performance is not meaningfully improved by introducing dependency structures into shrinkage priors. In most configurations, the distribution of ELPD differences is centered near zero, suggesting that dependency-aware priors perform comparably to their standard counterparts, at least on average. This pattern is particularly evident in the low dimensional setting, where even under high correlation and clearly structured signals, the inclusion of Ω offers no measurable advantage.

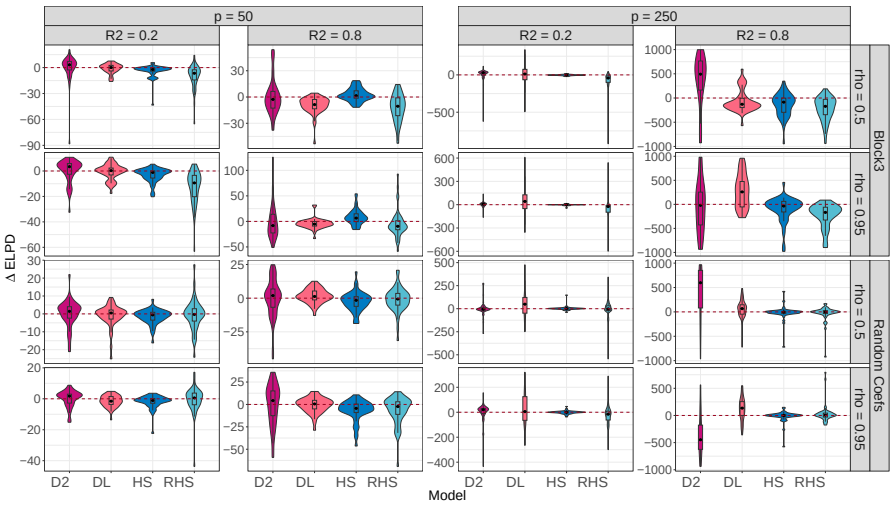


FIG 6. ***ΔELPD for each model under the BMA1 structure across simulation scenarios.*** *Model names represent the difference between the dependency-aware and standard versions, where positive values indicate improved predictive performance of the dependency-aware versions. We have omitted the BP prior due to high variability.*

Similarly, in high dimensional settings, the inclusion of dependency structures does not produce strong predictive improvements across priors and design scenarios. These results imply that standard shrinkage priors are already capable of achieving strong predictive performance and that additional modeling flexibility through dependence structures does not translate into predictive gains. However, a notable observation is that predictive accuracy remains stable across all conditions, indicating that dependency-aware priors do not harm generalization. While gains in prediction are limited, these priors remain useful when the focus shifts to parameter recovery, which we discuss below.

Figures 7 and 8 illustrate the impact of incorporating correlational information into the prior on parameter recovery under the fixed block coefficient setting, for the BMA1 and BAR1 structures, respectively. Figure 7 shows that incorporating dependency information consistently reduces RMSE for nonzero coefficients across all scenarios. Notably, improvements appear even in the most challenging setting: high dimensionality ($p = 250$), low signal ($R^2 = 0.2$), and strong correlation ($\rho = 0.95$). When the signal is stronger ($R^2 = 0.8$), and thus noise is reduced, dependency-aware shrinkage priors still yield substantial improvements in parameter recovery. These results support the hypothesis that when Ω accurately reflects the groupwise structure of the relevant predictors, dependency-aware priors facilitate more effective signal recovery. Figure 8 focuses on recovery of nonzero coefficients under the BAR1 structure. We limit the display to signal coefficients due to space constraints, but the results for all and zero coefficients mirror the patterns observed under BMA1. Contrary to

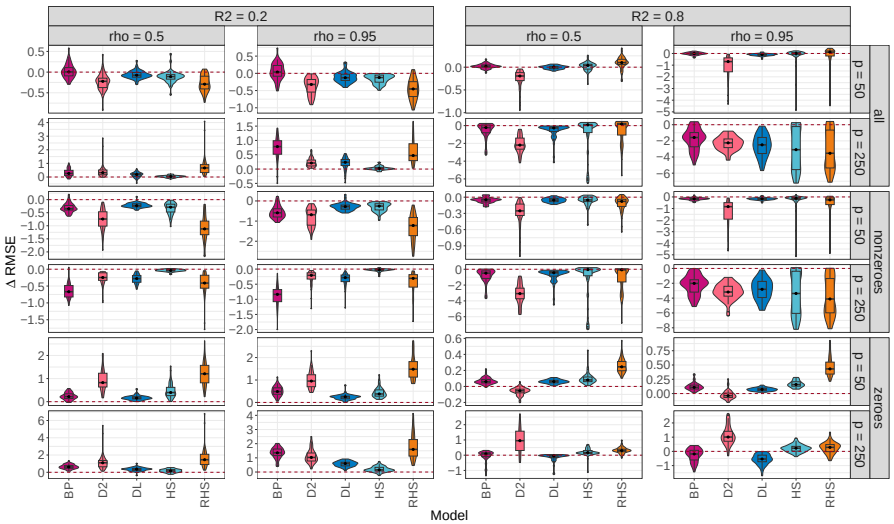


FIG 7. **Δ *RMSE for each model under the BMA1 structure with fixed block signals.*** *Rows correspond to all coefficients, nonzero (signal) coefficients, and zero (noise) coefficients. Model names represent the difference between the dependency-aware and standard versions. Negative values indicate improved parameter recovery of the dependency-aware versions. Dependency-aware priors show consistent gains for signal coefficients, while differences for noise coefficients are more modest.*

BMA1, however, recovery of nonzero coefficients for BAR1 is not improved, and in fact, RMSE worsens uniformly across conditions.

The difference in performance between BMA1 and BAR1 can be attributed to the structure of the prior correlation matrix Ω. While both scenarios assign strong signals to groups of correlated predictors, the implied Ω matrices differ in how they encode these relationships. In BAR1, Ω has a tridiagonal block structure, inducing local dependencies that connect only neighboring predictors within each block. In contrast, BMA1 produces blocks in which all predictors within a block are strongly correlated with one another, resulting in a dense (possibly high) correlation pattern. This richer within-block structure in Ω leads to stronger prior correlations among the nonzero coefficients, which enhances parameter recovery in the presence of structured signals. These results suggest that the effectiveness of dependency-aware priors depends not just on the presence of correlated signals, but on how exactly the signals are correlated.

3.2. Real world case studies

We evaluate how incorporating covariate dependence structures affects the predictive performance of shrinkage priors using two real high dimensional datasets. The first dataset, Cereal, contains starch content measurements for 15 samples, each with 145 infrared spectral predictors, and is provided by the `R` package `chemometrics` (Filzmoser, 2023). The second dataset, Eye, includes gene ex-

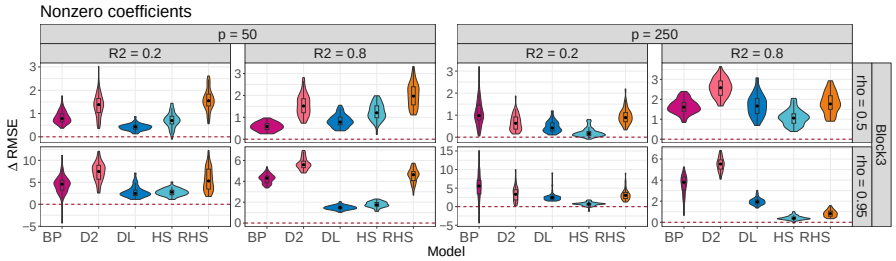


FIG 8. ***Δ RMSE for each model under the BMA1 structure with fixed block signals.*** *Only nonzero coefficients are shown. Model names represent the difference between the dependency-aware and standard versions, where negative values indicate improved parameter recovery of the dependency-aware versions. Improvements are limited and parameter recovery often worsens across conditions.*

pression levels for 200 genes across 120 samples obtained from microarray experiments on mammalian eye tissue (Scheetz et al., 2006) and is available in the `R` package `flare` (Li et al., 2024). Figure 9 displays histograms of the pairwise correlations among covariates.

To assess out-of-sample predictive performance, we use ELPD. As an independent test set is not available, we rely on exact leave-one-out cross-validation (LOO-CV) (Vehtari et al., 2016). For each fold, the ELPD is computed on the held-out observation and then aggregated across all folds to obtain an overall estimate. We compare the performance of the Horseshoe, Regularized Horseshoe, Dirichlet-Laplace, and R2D2 priors, along with their dependency-aware variants (Carvalho et al., 2010; Piironen and Vehtari, 2017; Bhattacharya et al., 2015).

We summarize the results in Table 1, which reports ELPD differences and standard deviations computed via pairwise comparisons against the best-performing model for each dataset. Negative values indicate lower predictive accuracy relative to the best model. Across real world correlation structures, a consistent pattern emerges: standard shrinkage priors often outperform their dependency-aware counterparts, even in settings with substantial predictor dependence (see Figure 9). While one might expect that explicitly modeling such correlations would enhance predictive performance, the empirical results suggest that standard priors already adapt well to complex structures without requiring explicit dependence modeling.

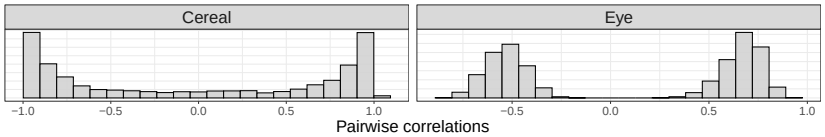


FIG 9. *Histograms of pairwise correlations among covariates across the datasets used in the case study.*

TABLE 1
Differences in ELPD and standard deviations for the datasets, computed via pairwise comparisons with the model having the highest ELPD (first column). The initial value is zero, and subsequent columns display negative values that indicate the difference with the best model. See Vehtari et al. (2016) for details on standard error calculations. Models ending in "O" denote dependency-aware variants and are highlighted in bold blue for emphasis.

Dataset								
Cereal	**DLO**	DL	D2	**D2O**	HS	RHS	**HSO**	**RHSO**
$(n, p) = (15, 145)$	0.0 (0.0)	-4.5 (2.9)	-4.9 (3.4)	-8.7 (5.6)	-17.7 (11.5)	-20.1 (12.5)	-20.7 (11.6)	-27.7 (13.4)
Eye	HS	D2	RHS	**HSO**	**RHSO**	**D2O**	DL	**DLO**
$(n, p) = (120, 200)$	0.0 (0.0)	-2.5 (2.3)	-2.6 (1.1)	-5.8 (1.5)	-10.0 (1.6)	-12.7 (2.3)	-14.2 (4.6)	-22.4 (5.2)

4. Discussion

We propose an extension to the continuous global-local shrinkage priors framework that allows for the inclusion of dependence structures via a correlation matrix Ω. Specifically, we move from the standard independent prior on each coefficient, $b_i \mid \lambda_i, \tau, \sigma \sim \mathcal{N}(0, \lambda_i^2 \tau^2 \sigma^2)$, to a multivariate prior $b \mid \sigma, \tau, \lambda \sim \mathcal{N}\left(0, \sigma^2 \tau^2 \mathrm{D}_\lambda \Omega \mathrm{D}_\lambda\right)$, where D_λ is a diagonal matrix containing the local scales λ_i, and Ω is a correlation matrix encoding dependencies among coefficients.

Importantly, our construction imposes no additional constraints on the distributions of the local scales λ_i, the global scale τ, or their hyperparameters. This preserves the full flexibility of the original shrinkage prior framework. It enables a multivariate generalization of continuous shrinkage priors such as the Horseshoe, without restricting their marginal heavy-tailed behavior or imposing constraints on finite-moment conditions, as was necessary in some earlier approaches (Griffin and Hoff, 2024). In the absence of knowledge about the true dependence structure, we propose estimating Ω using information from the design matrix X, following ideas akin to Zellner's g prior. When $p < n$, we use the empirical covariance matrix of X. In the high dimensional case ($p > n$), we adopt an automated shrinkage-based estimator that is computationally efficient and enables a fully automated specification of Ω.

Our simulation studies yield mixed results regarding the benefits of incorporating dependence structures. While the intuition that modeling correlations can improve performance is appealing, our results indicate that the benefits are scenario-specific. In particular, improvements in parameter recovery emerge when groups of highly correlated predictors align with the true signal structure. We attribute this to the regularization effect of Ω on the joint prior for b, which tends to mitigate excessive shrinkage of correlated signals. However, when the signal structure does not reflect the correlation pattern, such as in scenarios with randomly distributed nonzero coefficients, this alignment is lost, and the incorporation of Ω offers no meaningful advantage. Importantly, we find that predictive performance remains stable even when Ω does not match the signal, suggesting that dependency-aware priors can be safely employed without risk of degradation.

Our approach to incorporating dependency structures via Ω in the shrink-

age prior is only one possible strategy for modeling dependencies. Alternatives include placing a prior directly over Ω or requiring the user to specify it manually. Placing a prior over Ω results in a highly overparameterized model, with $p(p-1)/2$ additional parameters to estimate, increasing data requirements and computational burden. Manual specification shifts a heavy burden onto the user, requiring explicit modeling of $p(p-1)/2$ correlations, an impractical task in most real world applications. For shrinkage priors to remain practically useful, it is essential that their complexity remains manageable, ideally governed by a small number of hyperparameters, which is satisfied by our approach.

Our study pursued two specific objectives: First, to develop a streamlined, automated procedure for introducing dependence structures into shrinkage priors. Second, to evaluate whether doing so yields tangible practical benefits. The latter question is often overlooked in the literature, where methodological innovations are proposed without systematically investigating whether they meaningfully improve inference or prediction. Ultimately, while the incorporation of dependence structures into shrinkage priors is theoretically appealing, our results suggest that in many practical settings, the added complexity is difficult to justify. Standard shrinkage priors, with their simplicity, robustness, and strong predictive performance, often provide a better balance between model flexibility and computational efficiency. We therefore recommend that practitioners prioritize simpler formulations unless there is strong prior evidence of structured correlations among predictors. In short, dependence modeling should be seen not as a default strategy, but as a targeted tool—useful when needed, but costly when applied indiscriminately.

Acknowledgments

Funded by Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) Project 500663361. We acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by DFG Project 271512359. Paul-Christian Bürkner further acknowledges support of the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) via the Collaborative Research Center 391 (Spatio-Temporal Statistics for the Transition of Energy and Transport) – 520388526.

The authors would like to thank Yuexi Wang, Luna Fazio, and David Kohns for their thoughtful comments and discussion on earlier versions of the manuscript.

5. Appendix

5.1. Shrinkage priors

We consider the following shrinkage priors, with default hyperparameter settings detailed below.

Following the recommendations of Gelman (2006), we place a Half Student-t prior on the error scale parameter σ, with ν degrees of freedom and scale parameter η. We set $\eta \approx \text{sd}(y)$, as both the prior mean and variance are proportional to η.

5.1.1. Beta Prime

The Beta Prime (BP) prior is specified via:

$$b_j \mid \lambda_j, \sigma \sim \mathcal{N}\left(0, \lambda_j^2 \sigma^2\right), \ \lambda_j^2 \sim \text{BetaPrime}(\alpha, \beta), \ \sigma \sim p(\sigma)$$

where $\text{BetaPrime}(\alpha, \beta)$ has density proportional to $\lambda^{\alpha-1}(1+\lambda)^{-\alpha-\beta}$ (Bai and Ghosh, 2019, 2021). The suggested default choices are $\alpha \leq 0.5$ to generate unbounded marginal distributions for b. We set $\alpha \sim \text{Gamma}(1, 2)$ and $\beta = 1$.

5.1.2. Dirichlet Laplace

The Dirichlet Laplace (DL) prior induces global-local shrinkage through the following hierarchy:

$$\begin{aligned} b_j \mid \psi_j, \phi, \tau, \sigma &\sim \mathcal{N}\left(0, \psi_j(\phi_j\tau)^2\sigma^2\right), \\ \psi_j &\sim \text{Exp}\left(1/2\right), \\ \phi = (\phi_1, \ldots, \phi_p) &\sim \text{Dirichlet}(a_\pi, \ldots, a_\pi), \\ \tau &\sim \text{Gamma}(na_\pi, 1/2) \\ \sigma &\sim p(\sigma) \end{aligned}$$

where $\text{Exp}(1/2)$ denotes the exponential distribution with rate $1/2$. We set $a_\pi = 0.5$ as suggested in the Simulation Study by Bhattacharya et al. (2015).

5.1.3. Horseshoe

The Horseshoe prior is defined hierarchically as:

$$b_j \mid \lambda_j, \tau, \sigma \sim \mathcal{N}\left(0, \sigma^2\lambda_j^2\tau^2\right), \ \lambda_j \sim \mathcal{C}^+(0,1), \ \tau \sim \mathcal{C}^+(0,1), \ \sigma \sim p(\sigma)$$

where $\mathcal{C}^+(0,1)$ denotes the half-Cauchy distribution with unit scale (Carvalho et al., 2010). Notice that the Horseshoe prior does not possess hyperparameters other than the one present in the distribution of σ.

5.1.4. Normal-Gamma

The Normal Gamma prior proposed by Brown and Griffin (2010) is specified as follows:

$$\begin{aligned}
\beta_i \mid \psi_i, \sigma &\sim \mathcal{N}(0, \sigma^2 \psi_i^2), \\
\psi_i^2 \mid \lambda, \gamma &\sim \text{Gamma}\left(\lambda, \frac{1}{2\gamma^2}\right), \\
\nu &= 2\lambda\gamma^2, \\
\nu \mid \lambda &\sim \text{InverseGamma}(2, M), \\
\lambda &\sim \text{Exponential}(1).
\end{aligned}$$

Brown and Griffin (2010) argue that centering the prior for λ around 1 introduces variability around the Bayesian LASSO prior. In this specification, ν has expectation M, which serves as an informed guess for the squared ℓ_2-norm of b. We follow their same approach and set

$$M = \frac{1}{p} \sum_{i=1}^{p} \hat{b}_i^2,$$

where $\hat{b}$ is the maximum likelihood estimator (MLE) of b when X has full rank, and the minimum-length least squares estimator when $p > n$ (Brown and Brown, 1994). Setting $\lambda = 1/2$, $M = 1$ and $1/\nu \sim \text{Gamma}(1/2, 1)$ results in the Horseshoe prior (Griffin and Brown, 2021).

5.1.5. R2D2

The R2D2 prior places a distribution on the proportion of explained variance R^2, and allocates the total prior variance ω^2 among coefficients using a Dirichlet distribution Zhang et al. (2020). Specifically, a Beta prior is assigned to R^2, which induces a Beta Prime prior on ω^2 with parameters $a_1, a_2 > 0$.

The total variance ω^2 is decomposed across coefficients using proportions $\phi = (\phi_1, \ldots, \phi_p)$, where $\phi \sim \text{Dirichlet}(\alpha)$ and $\lambda_i^2 = \phi_i \omega^2$. Originally, Zhang et al. (2020) specified a Double Exponential prior for the coefficients, while Aguilar and Bürkner (2023) proposed a Normal prior. Following the recommendations of Aguilar and Bürkner (2023), we adopt a Normal prior specification and set the hyperparameters $\alpha = (a_\pi, ..., a_\pi)$ with $a_\pi = 0.25$, $a_2 = 0.5$ to ensure unbounded marginal distributions and heavy-tailed behavior. Aguilar and Bürkner (2025) show that other decompositions of variance are available, however we have only considered the Dirichlet in our experiments.

The full hierarchical structure of the R2D2 prior is:

$$\begin{aligned}
b_j \mid \lambda_j^2, \sigma^2 &\sim \mathcal{N}\left(0,\, \sigma^2 \lambda_j^2\right), \\
\lambda_j^2 &= \phi_j \omega^2, \\
\phi &\sim \text{Dirichlet}(a_\pi, \ldots, a_\pi), \\
\omega^2 &\sim \text{BetaPrime}(a_1, a_2)\, \sigma \sim p(\sigma).
\end{aligned}$$

5.2. Bounds

In this subsection we prove the bounds presented in the main text. Our objective is to bound the spectral norm of $Q_\Omega^{-1} - Q_I^{-1}$, where

$$\begin{aligned} Q_\Omega &= X'X + D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}, \\ Q_I &= X'X + D_\lambda^{-2}. \end{aligned} \tag{26}$$

In the following let $\lambda_1 = \max_i \lambda_i, \lambda_p = \min_i \lambda_i, \nu_1 = \lambda_{\max}(X'X), \nu_p = \lambda_{\min}(X'X), \omega_1 = \lambda_{\max}(\Omega), \omega_p = \lambda_{\min}(\Omega)$ where $\lambda_{\max}(A)$ and $\lambda_{\min}(A)$ denote the maximum and minimum eigenvalues of A respectively.

We make use of the following:

- Let $A \in \mathbb{R}^{p\times p}$ be a real symmetric matrix and $x \in \mathbb{R}^p$ a nonzero vector. Then its spectral norm is given by

$$\|A\|_2 = \max_{x\neq 0} \frac{\|Ax\|_2}{\|x\|_2} = \max_i |\lambda_i(A)|, \tag{27}$$

  where $\lambda_i(A)$ denotes the i-th eigenvalue of A. In particular, if A is positive semidefinite, the spectral norm equals its largest eigenvalue.
- The resolvent identity, which states that for invertible matrices A and B:

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}. \tag{28}$$

- The submultiplicative property of the spectral norm:

$$\|AB\|_2 \leq \|A\|_2\|B\|_2, \tag{29}$$

- If A and C are invertible then:

$$\|ABC\|_2 \geq \|A^{-1}\|_2^{-1}\|B\|_2 \cdot \|C^{-1}\|_2^{-1}. \tag{30}$$

- Rayleigh characterization. Let $A \in \mathbb{R}^{p\times p}$ be a real symmetric matrix. Then its smallest and largest eigenvalues can be characterized as (Anderson, 2003)

$$\lambda_{\min}(A) = \min_{\|x\|_2=1} x'Ax, \qquad \lambda_{\max}(A) = \max_{\|x\|_2=1} x'Ax. \tag{31}$$

  This implies that $\lambda_{\min}(A) \leq x'Ax \leq \lambda_{\max}(A), \forall x \in \{x | \|x\|_2 = 1\}$.
- Weyl's inequality for symmetric matrices A and B (Weyl, 1912):

$$\begin{aligned} \lambda_{\min}(A + B) &\geq \lambda_{\min}(A) + \lambda_{\min}(B), \\ \lambda_{\max}(A + B) &\leq \lambda_{\max}(A) + \lambda_{\max}(B). \end{aligned} \tag{32}$$

We begin by proving the upper bound. An application of the resolvent identity yields

$$Q_\Omega^{-1} - Q_I^{-1} = Q_\Omega^{-1}(Q_I - Q_\Omega)Q_I^{-1} = Q_\Omega^{-1}D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}Q_I^{-1}. \tag{33}$$

Taking spectral norms and applying the submultiplicative property:

$$\|Q_\Omega^{-1} - Q_I^{-1}\|_2 \leq \|Q_\Omega^{-1}\|_2 \cdot \|D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}\|_2 \cdot \|Q_I^{-1}\|_2.$$

We now proceed to bound each term. First, applying the submultiplicative property to the norm in the middle of the right hand, we have

$$\|D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}\|_2 \leq \|D_\lambda^{-1}\|_2^2 \|\Omega^{-1} - I\|_2 = \frac{1}{\lambda_p^2}\|\Omega^{-1} - I\|_2. \tag{34}$$

Next, we bound the spectral norms of Q_Ω^{-1} and Q_I^{-1}. Since Q_Ω and Q_I are symmetric positive definite, the spectral norms of Q_Ω^{-1} and Q_I^{-1} are given by the reciprocals of their smallest eigenvalues of Q_Ω and Q_I respectively. By Weyl's inequality:

$$\begin{aligned}\lambda_{\min}(Q_\Omega) = \lambda_{\min}\left(X'X + D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}\right) &\geq \lambda_{\min}(X'X) + \lambda_{\min}(D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}) \\ &\geq \nu_p + \frac{1}{\lambda_1^2\omega_1}.\end{aligned} \tag{35}$$

In the last line we have used

$$\lambda_{\min}(D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}) = \frac{1}{\lambda_{\max}(D_\lambda\Omega D_\lambda)} \geq \frac{1}{\lambda_{\max}(D_\lambda^2)\lambda_{\max}(\Omega)} = \frac{1}{\lambda_1^2\omega_1},$$

which follows from the submultiplicative property of the spectral norm.

Using the same argument we obtain

$$\lambda_{\min}(Q_I) \geq \nu_p + \frac{1}{\lambda_1^2}.$$

Therefore

$$\|Q_\Omega^{-1}\|_2 \leq \left(\nu_p + \frac{1}{\lambda_1^2\omega_1}\right)^{-1}, \quad \|Q_I^{-1}\|_2 \leq \left(\nu_p + \frac{1}{\lambda_1^2}\right)^{-1}. \tag{36}$$

Combining Equation (34) and Equation (36) yields

$$\|Q_\Omega^{-1} - Q_I^{-1}\|_2 \leq \frac{\|\Omega^{-1} - I\|_2}{\lambda_p^2\left(\nu_p + \frac{1}{\lambda_1^2\omega_1}\right)\left(\nu_p + \frac{1}{\lambda_1^2}\right)} \tag{37}$$

We proceed to prove the lower bound. From the identity:

$$Q_\Omega^{-1} - Q_I^{-1} = Q_\Omega^{-1}D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}Q_I^{-1},$$

we apply the reverse inequality for products of invertible matrices shown in Equation (30) to get

$$\|Q_\Omega^{-1} - Q_I^{-1}\|_2 \geq \|Q_\Omega\|_2^{-1}\|D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}\|_2\|Q_I\|_2^{-1}. \tag{38}$$

To lower bound $\|Q_\Omega\|_2^{-1}$ and $\|Q_I\|_2^{-1}$ we make use of Weyl's inequality on $\|Q_\Omega\|_2$ and $\|Q_I\|_2$:

$$\begin{aligned}\lambda_{\max}(Q_\Omega) = \lambda_{\max}\left(X'X + D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}\right) &\leq \lambda_{\max}(X'X) + \lambda_{\max}(D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}) \\ &\leq \nu_1 + \frac{1}{\lambda_p^2\omega_p}.\end{aligned} \tag{39}$$

In the last line we have used that

$$\lambda_{\max}(D_\lambda^{-1}\Omega^{-1}D_\lambda^{-1}) = \frac{1}{\lambda_{\min}(D_\lambda\Omega D_\lambda)} \leq \frac{1}{\lambda_{\min}(D_\lambda^2)\lambda_{\min}(\Omega)} = \frac{1}{\lambda_p^2\omega_p}.$$

This statement follows from the Rayleigh characterization shown in 31. Given that $\lambda_{\min}(D_\lambda\Omega D_\lambda) = \min_{\|x\|_2=1} x'D_\lambda\Omega D_\lambda x$ and letting $y = D_\lambda x$ leads to $\|y\|^2 = y'y = x'D_\lambda^2 x$. This implies that $\lambda_{\min}(D_\lambda^2) \leq \|y\|^2 \leq \lambda_{\max}(D_\lambda^2)$. Therefore

$$x'D_\lambda\Omega D_\lambda x \geq y'\Omega y \geq \lambda_{\min}(\Omega)y'y.$$

Taking the minimum yields

$$\lambda_{\min}(D_\lambda\Omega D_\lambda) \geq \lambda_{\min}(\Omega)\lambda_{\min}(D_\lambda^2) = \omega_p\lambda_p^2.$$

A similar argument produces

$$\lambda_{\max}(Q_I) \leq \nu_1 + \frac{1}{\lambda_p^2}. \tag{40}$$

Therefore

$$\|Q_\Omega\|_2 \leq \nu_1 + \frac{1}{\lambda_p^2\omega_p}, \qquad \|Q_I\|_2 \leq \nu_1 + \frac{1}{\lambda_p^2}, \tag{41}$$

We proceed to bound $\|D_\lambda^{-1}(\Omega^{-1}-I)D_\lambda^{-1}\|_2$ in line 38. By applying the reverse inequality for products of invertible matrices we obtain

$$\|D_\lambda^{-1}(\Omega^{-1} - I)D_\lambda^{-1}\|_2 \geq \frac{1}{\lambda_1^2} \cdot \|\Omega^{-1} - I\|_2, \tag{42}$$

Combining Equation (41) and Equation (42) results in

$$\|Q_\Omega^{-1} - Q_I^{-1}\|_2 \geq \frac{\|\Omega^{-1} - I\|_2}{\lambda_1^2\left(\nu_1 + \frac{1}{\lambda_p^2\omega_p}\right)\left(\nu_1 + \frac{1}{\lambda_p^2}\right)}$$

Remark (explicit form of $\|\Omega^{-1} - I\|_2$). Since Ω is symmetric positive definite, it admits an eigendecomposition $\Omega = U \operatorname{diag}(\omega_1, \ldots, \omega_p)\, U'$ with $0 < \omega_p \leq \cdots \leq \omega_1$. Hence

$$\Omega^{-1} - I = U \operatorname{diag}\left(\frac{1}{\omega_1} - 1, \ldots, \frac{1}{\omega_p} - 1\right) U',$$

so the eigenvalues of $\Omega^{-1} - I$ are $\{1/\omega_i - 1\}_{i=1}^p$. Because $\Omega^{-1} - I$ is symmetric, its spectral norm equals the largest absolute eigenvalue, and therefore

$$\|\Omega^{-1} - I\|_2 = \max_i \left|\frac{1}{\omega_i} - 1\right| = \max\left\{\left|\frac{1}{\omega_p} - 1\right|,\ \left|\frac{1}{\omega_1} - 1\right|\right\}.$$

Since Ω is a correlation matrix, $\omega_p \leq 1 \leq \omega_1$ and this expression reduces to

$$\|\Omega^{-1} - I\|_2 = \max\left\{\frac{1}{\omega_p} - 1,\ 1 - \frac{1}{\omega_1}\right\}.$$

5.3. Concentration properties

Proposition 1 (Joint spike and joint tail mass under a common-scale special case)**.** *Let $p \geq 2$ and let Ω be a positive definite correlation matrix in $\mathbb{R}^{p\times p}$. Let $\lambda \sim \mathrm{C}^+(0,1)$ be independent of $z \sim \mathcal{N}(0,\Omega)$ and define $b = \lambda z$. Write $p_\Omega(b)$ for the joint density of b, and let $p(\cdot)$ denote the common one-dimensional marginal density of each b_i. Define $p_I(b) = \prod_{i=1}^p p(b_i)$. Then the following hold.*

(i) As $b \to 0$,

$$p_\Omega(b) \asymp \left(b'\Omega^{-1}b\right)^{-(p-1)/2}, \tag{43}$$

and in particular $p_\Omega(b)/p_I(b) \to \infty$.

(ii) Let $A = \{x \in \mathbb{R}^p : \min_{1\leq i\leq p} |x_i| > 1\}$. Then

$$P(b \in tA) \sim \frac{2}{\pi}\, \frac{\mathbb{E}\left[\min_{1\leq i\leq p} |z_i|\right]}{t} \qquad \text{as } t \to \infty, \tag{44}$$

while under $p(\cdot)$ one has $P_I(b \in tA) \asymp t^{-p}$. Consequently, $P(b \in tA)/P_I(b \in tA) \to \infty$.

Proof. First claim. Write $Q(b) = b'\Omega^{-1}b$. Conditionally on λ, we have $b \mid \lambda \sim \mathcal{N}(0, \lambda^2\Omega)$. Since $\lambda \sim \mathrm{C}^+(0,1)$ has density $f_\lambda(\lambda) = \frac{2}{\pi}(1+\lambda^2)^{-1}$ on $(0,\infty)$, the joint density of b admits the scale-mixture representation

$$p_\Omega(b) = c_\Omega \int_0^\infty \lambda^{-p} \exp\left(-\frac{Q(b)}{2\lambda^2}\right) \frac{1}{1+\lambda^2}\, d\lambda, \tag{45}$$

where $c_\Omega = \frac{2}{\pi}(2\pi)^{-p/2}|\Omega|^{-1/2}$. We decompose the integral at $\lambda = 1$. For $\lambda \geq 1$ the integrand is bounded by $\lambda^{-p}(1+\lambda^2)^{-1}$, because the exponential factor is at most one. Moreover, for $\lambda \geq 1$ we have $(1+\lambda^2)^{-1} \leq \lambda^{-2}$, so $\lambda^{-p}(1+\lambda^2)^{-1} \leq$

$\lambda^{-(p+2)}$, which is integrable on $(1, \infty)$. It follows that the contribution of the region $(1, \infty)$ is finite and bounded uniformly over b.

The singular behavior as $b \to 0$ comes from the region $\lambda \in (0, 1)$. Define

$$J_p(Q) = \int_0^1 \lambda^{-p} \exp\left(-\frac{Q}{2\lambda^2}\right) \frac{1}{1+\lambda^2}\, d\lambda \qquad \text{and} \qquad I_p(Q) = \int_0^1 \lambda^{-p} \exp\left(-\frac{Q}{2\lambda^2}\right) d\lambda.$$

On $(0, 1)$ we have $1 \leq 1 + \lambda^2 \leq 2$, hence $\frac{1}{2} \leq (1+\lambda^2)^{-1} \leq 1$, and therefore

$$\frac{1}{2} I_p(Q) \;\leq\; J_p(Q) \;\leq\; I_p(Q) \qquad \text{for all } Q \geq 0. \tag{46}$$

Consequently, $J_p(Q) \asymp I_p(Q)$ as $Q \downarrow 0$, so it suffices to obtain the asymptotic form of $I_p(Q)$. To evaluate $I_p(Q)$, substitute $t = Q/(2\lambda^2)$, so that $\lambda = \sqrt{Q/(2t)}$ and $d\lambda = -\frac{1}{2}\sqrt{Q/2}\, t^{-3/2}\, dt$. This gives

$$I_p(Q) = \frac{1}{2}\left(\frac{Q}{2}\right)^{-(p-1)/2} \int_{Q/2}^{\infty} t^{(p-3)/2} e^{-t}\, dt.$$

When $p \geq 2$, the exponent $(p-3)/2 > -1$, and the integral converges to $\Gamma\left(\frac{p-1}{2}\right)$ as $Q \downarrow 0$. Hence,

$$I_p(Q) \sim \frac{1}{2}\left(\frac{Q}{2}\right)^{-(p-1)/2} \Gamma\left(\frac{p-1}{2}\right), \qquad Q \downarrow 0.$$

Combining this with (46) shows that the same rate holds for $J_p(Q)$. Returning to Equation (45) and recalling that the contribution from $(1, \infty)$ is uniformly bounded, we conclude that, as $b \to 0$,

$$p_\Omega(b) \asymp Q(b)^{-(p-1)/2} = \left(b'\Omega^{-1}b\right)^{-(p-1)/2}.$$

By positive definiteness of Ω^{-1}, let $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and largest eigenvalues of Ω^{-1}. Then for all $b \in \mathbb{R}^p$, using Rayleigh characterization we have

$$\lambda_{\min} \|b\|^2 \;\leq\; b'\Omega^{-1}b \;\leq\; \lambda_{\max} \|b\|^2,$$

so $Q(b) = b'\Omega^{-1}b \asymp \|b\|^2$ as $b \to 0$. This implies

$$Q(b)^{-(p-1)/2} \asymp \|b\|^{-(p-1)} \qquad (b \to 0).$$

Because Ω is a correlation matrix, each component satisfies $b_i = \lambda z_i$ with $z_i \sim \mathcal{N}(0, 1)$. Thus each marginal density is the standard horseshoe marginal, which behaves around the origin as (Carvalho et al., 2009; Zhang et al., 2020; Aguilar and Bürkner, 2023)

$$p(b_i) = \mathcal{O}\big(\log(1/|b_i|)\big) \qquad \text{as } b_i \to 0.$$

Therefore,

$$p_I(b) = \prod_{i=1}^{p} p(b_i) = \mathcal{O}\left(\prod_{i=1}^{p} \log\left(\frac{1}{|b_i|}\right)\right) \qquad \text{as } b \to 0.$$

Since $|b_i| \leq \|b\|$ we have $\log(1/|b_i|) \leq \log(1/\|b\|)$ for $\|b\|$ small, which yields

$$p_I(b) = O\left(\log\left(\frac{1}{\|b\|}\right)^p\right).$$

Finally, any polynomial rate dominates any power of a logarithm, in the sense that for $\alpha > 0$ and $m \geq 1$,

$$\lim_{r \downarrow 0} r^\alpha \big(\log(1/r)\big)^m = 0.$$

With $r = \|b\|$, $\alpha = p - 1$, and $m = p$, this implies $\|b\|^{-(p-1)}/\log(1/\|b\|)^p \to \infty$. Combining the bounds gives $p_\Omega(b)/p_I(b) \to \infty$ as $b \to 0$, which proves the first claim.

Second claim.

Consider the set that represents that every coordinate of a vector x is at least 1:

$$A = \left\{x \in \mathbb{R}^p : \min_{1 \leq i \leq p} |x_i| > 1\right\}.$$

Then $b \in tA$ if and only if $\min_i |b_i| > t$. Under the global-scale representation we have $b = \lambda z$, therefore $\min_i |b_i| = \min_i |\lambda z_i| = \lambda \min_i |z_i|$. Writing

$$m(z) = \min_{1 \leq i \leq p} |z_i|,$$

gives the equivalence

$$\{b \in tA\} = \{\lambda m(z) > t\} = \left\{\lambda > \frac{t}{m(z)}\right\}.$$

Conditioning on z yields

$$P(b \in tA) = \mathbb{E}\left[P\left(\lambda > \frac{t}{m(z)} \;\middle|\; z\right)\right] = \mathbb{E}\left[\bar{F}_\lambda\left(\frac{t}{m(z)}\right)\right], \tag{47}$$

where $\bar{F}_\lambda(u) = P(\lambda > u)$ is the survival function of λ. For the half-Cauchy scale, the tail satisfies

$$\bar{F}_\lambda(u) \sim \frac{2}{\pi}\frac{1}{u} \qquad \text{as } u \to \infty.$$

Since $m(z) \in (0, \infty)$ and $\mathbb{E}[m(z)] < \infty$ for a multivariate normal z, we can apply dominated convergence to Equation (47) to obtain

$$P(b \in tA) \sim \frac{2}{\pi}\frac{\mathbb{E}[m(z)]}{t} \qquad \text{as } t \to \infty.$$

For the product construction, the coordinates are independent and each marginal satisfies

$$P(|b_i| > t) \asymp t^{-1}.$$

Therefore

$$P_I(b \in tA) = P\Big(\min_i |b_i| > t\Big) = \prod_{i=1}^{p} P(|b_i| > t) \asymp t^{-p}.$$

Consequently,

$$\frac{P(b \in tA)}{P_I(b \in tA)} \to \infty \qquad \text{as } t \to \infty,$$

which proves the second claim. □

5.4. Correlation Matrices

We have used the following correlation matrix structures in the main text. For further details on these and related constructions, see Rue and Held (2005) for details.

5.4.1. Autoregressive of Order 1

The autoregressive correlation matrix of order 1, AR(1), assumes that the correlation between two observations decays exponentially with their distance. Specifically, for observations i and j, the (i, j)-th entry of the matrix is given by

$$\text{Corr}(i, j) = \rho^{|i-j|}, \tag{48}$$

where $\rho \in (-1, 1)$ is the autocorrelation parameter.

5.4.2. Moving Average of order 1

The moving average correlation matrix of order 1 MA(1) captures correlations between immediate neighbors. The (i, j)-th entry of the matrix is

$$\text{Corr}(i, j) = \begin{cases} 1 & \text{if } i = j, \\ \rho & \text{if } |i - j| = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho \in (-1, 1)$ measures the strength of correlation between adjacent observations.

5.4.3. Moving Average of order 2

The moving average correlation matrix of order 2 MA(2) extends the MA(1) structure to also include second neighbors. The (i, j)-th entry of the matrix is

$$\text{Corr}(i, j) = \begin{cases} 1 & \text{if } i = j, \\ \rho_1 & \text{if } |i - j| = 1, \\ \rho_2 & \text{if } |i - j| = 2, \\ 0 & \text{otherwise,} \end{cases}$$

where $\rho_1, \rho_2 \in (-1, 1)$ and control the correlation between first and second neighbors, respectively. Additionally it is required that $\rho_1 \pm \rho_2 < 1$. To parametrize the MA(2) in terms of only $\rho \in (-1, 1)$ we have set $\rho_1 = \rho$ and $\rho_2 = (1 - \rho)\rho$.

5.5. Ledoit-Wolf estimator

Ledoit and Wolf (2004b) propose the following linear shrinkage estimator

$$\Sigma^* = \varphi_1 I + \varphi_2 S = \frac{\beta^2}{\delta^2}\mu I + \frac{\alpha^2}{\delta^2} S. \tag{49}$$

where $\delta^2 = \alpha^2 + \beta^2$ and Σ^* minimizes the expected quadratic loss $\mathbb{E}\left\|\tilde{\Sigma} - \Sigma_X\right\|_F^2$ subject to $\tilde{\Sigma} = \varphi_1 I + \varphi_2 S$, with respect to nonrandom coefficients φ_1, φ_2. Here, $\|\cdot\|_F$ denotes the scaled Frobenius norm, i.e., $\|A\|_F = \mathrm{tr}(A'A)/p$. The shrinkage parameters used in the estimator are computed as follows:

$$\mu = \frac{\mathrm{tr}(S)}{p}, \quad \alpha^2 = \|\Sigma_X - \mu I\|_F^2, \quad \beta^2 = \mathbb{E}\|S - \Sigma_X\|_F^2. \tag{50}$$

Since Σ_X is unknown, consistent estimators for μ, α^2, and β^2 must be used to construct a consistent estimator S^* of Σ^*.

Proposition 2. *Let $X \in \mathbb{R}^{n \times p}$ be a design matrix, and define*

$$m_n = \frac{\mathrm{tr}(S)}{p}, \quad d_n^2 = \|S - m_n I\|_F^2,$$
$$b_n^2 = \min\left(\frac{1}{n^2}\sum_{k=1}^{n}\|x_{k\cdot}x_{k\cdot}' - S\|_F^2,\ d_n^2\right), \quad a_n^2 = d_n^2 - b_n^2,$$

where $x_{k\cdot}$ denotes the k-th row of X, and S is the sample covariance matrix. Then:

1. *m_n, d_n^2, and b_n^2 are consistent (in quadratic mean) estimators of μ, δ^2, and β^2, respectively.*
2. *The resulting linear shrinkage estimator*

$$S^* = \frac{b_n^2}{d_n^2} m_n I + \frac{a_n^2}{d_n^2} S \tag{51}$$

is a consistent estimator of the Ledoit–Wolf shrinkage estimator Σ^*.

See Ledoit and Wolf (2004b) for proof and further details.

5.6. Simulations: Further results

5.6.1. Coverage

Tables 2 and 3 summarize the coverage performance of various shrinkage priors under two correlation levels ($\rho = 0.5$ and $\rho = 0.95$), with fixed $p = 250$ and

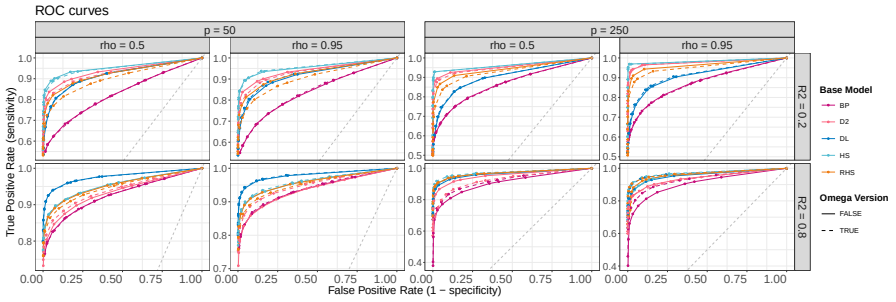


FIG 10. ***ROC curves for the BMA1 with fixed block signals*** *($p \in \{50, 250\}$, $R^2 \in \{0.2, 0.8\}$). We compare standard shrinkage priors (solid lines) and their dependency-aware counterparts (dashed lines). Models with an "O" suffix incorporate the BMA1 structure to induce dependency-aware shrinkage. Performance gains are most visible at high ρ and high R^2, where structured shrinkage improves signal recovery.*

high signal strength ($R^2 = 0.8$) when using 95% posterior credibility intervals using the BMA1 and BAR1 structures respectively. Models with an "O" suffix correspond to dependency-aware priors. The high-R^2 setting ensures that signal is present and identifiable, making this a meaningful context to evaluate how different priors handle multicollinearity. When R^2 is low, most of the variation in the response remains unexplained, making it difficult to discern meaningful signals from noise regardless of the prior.

Regarding the results of structure BMA1 shown in Table 2, all models maintain high coverage and specificity across settings. The most striking differences emerge in sensitivity and nonzero coverage, particularly as ρ increases. While base priors suffer sharp declines in power at high ρ (for instance the HS drops from 0.729 to 0.352 while HSO moves from 0.921 to 0.906), dependency-aware priors remain robust, consistently recovering signals with high probability. This echoes the results of parameter recovery that we show in the Methods Section of the main paper.

In contrast to the BMA1 structure, where dependency-aware priors delivered clear improvements under high collinearity, the BAR1 Fixed design proves more challenging. At $\rho = 0.5$, all models achieve high coverage and specificity, but sensitivity remains modest, generally between 0.2 and 0.5. Notably, the dependency-aware variants often underperform their unstructured counterparts, suggesting that incorporating the BAR1-based precision matrix can overshrink when the signal is not well aligned with the imposed structure. As ρ increases to 0.95, sensitivity deteriorates across all models, falling below 0.15. Even dependency-aware priors, which excelled under BMA1, fail to recover meaningful signal here.

Figures 10 and 11 show ROC curves under the BMA1 and BAR1 structures, respectively, across varying levels of correlation ($\rho \in 0.5, 0.95$), dimensionality ($p \in 50, 250$), and signal strength ($R^2 \in 0.2, 0.8$). As before, each standard shrinkage prior is plotted alongside its dependency-aware version (dashed line of the same color).

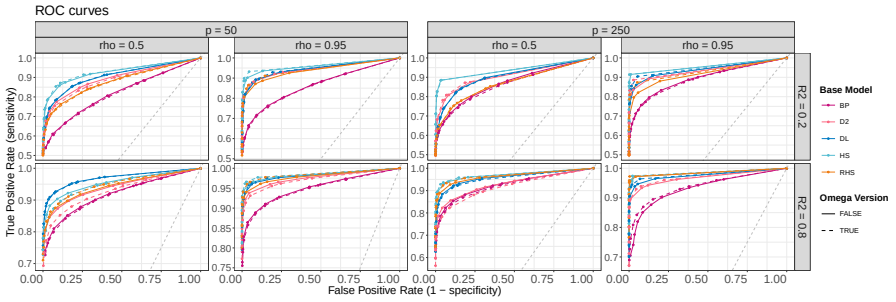


FIG 11. ***ROC curves for the BAR1 with fixed block signals*** *(*$p \in \{50, 250\}$*,* $R^2 \in \{0.2, 0.8\}$*). We compare standard shrinkage priors (solid lines) and their dependency-aware counterparts (dashed lines). Models with an "O" suffix incorporate the BAR1 structure to induce dependency-aware shrinkage. As a result, performance is comparable or slightly degraded relative to the standard versions.*

Under the BMA1 structure, dependency-aware priors consistently outperform or match their unstructured counterparts, especially at high ρ and high R^2. In these settings, dependency-aware shrinkage priors improve true positive rates without sacrificing specificity, resulting in ROC curves that dominate or closely track above their base versions. In contrast, under the BAR1 structure, the dependency-aware variants perform similarly or slightly worse than the standard priors. The ROC curves largely overlap, and in some cases the dependency-aware versions lag behind. This reinforces the earlier observation that the utility of structured priors is highly design dependent: they are most effective when the imposed dependence structure is aligned with the true generative mechanism (as in BMA1), but offer limited or no benefit when the structure is mismatched (as in BAR1).

TABLE 2

Coverage metrics under the BMA1 structure with fixed block signals *($p = 250$,* ***$R^2 = 0.8$****). We compare shrinkage priors across two correlation levels ($\rho = 0.5$ and $\rho = 0.95$). Models with an "O" suffix incorporate the BMA1 structure to induce dependency-aware shrinkage.*

Model ID	Coverage	Specificity	Sensitivity (Power)	Avg. CI Width	Coverage Zero	Coverage Nonzero
Condition: $\rho = 0.5$						
BP	0.984	0.996	0.831	1.032	0.996	0.696
BPO	0.989	0.997	0.990	1.028	0.997	0.794
D2	0.961	1.000	0.067	0.972	1.000	0.029
D2O	0.971	0.998	0.733	1.156	0.998	0.321
DL	0.989	0.999	0.860	0.869	0.999	0.733
DLO	0.992	0.999	0.988	0.819	0.999	0.817
HS	0.987	1.000	0.729	0.603	1.000	0.688
HSO	0.994	1.000	0.921	0.669	1.000	0.850
RHS	0.990	1.000	0.721	0.694	1.000	0.756
RHSO	0.993	1.000	0.942	0.758	1.000	0.825
Condition: $\rho = 0.95$						
BP	0.974	0.995	0.569	0.916	0.995	0.463
BPO	0.984	0.997	0.983	0.891	0.997	0.656
D2	0.960	1.000	0.065	0.740	1.000	0.000
D2O	0.967	1.000	0.865	1.008	1.000	0.181
DL	0.976	0.999	0.494	0.787	0.999	0.419
DLO	0.989	1.000	0.967	0.674	1.000	0.735
HS	0.975	1.000	0.352	0.457	1.000	0.371
HSO	0.991	1.000	0.906	0.550	1.000	0.781
RHS	0.967	0.993	0.310	0.557	0.993	0.346
RHSO	0.991	1.000	0.940	0.636	1.000	0.769

TABLE 3

Coverage metrics under the BAR1 structure with fixed block signals *($p = 250$,* ***$R^2 = 0.8$****). We compare shrinkage priors across two correlation levels ($\rho = 0.5$ and $\rho = 0.95$). Models with an "O" suffix incorporate the BMA1 structure to induce dependency-aware shrinkage.*

Model ID	Coverage	Specificity	Sensitivity (Power)	Avg. CI Width	Coverage Zero	Coverage Nonzero
Condition: $\rho = 0.5$						
BP	0.987	0.995	0.506	2.358	0.995	0.802
BPO	0.986	0.996	0.281	2.645	0.996	0.746
D2	0.996	1.000	0.492	2.385	1.000	0.915
D2O	0.993	1.000	0.158	2.869	1.000	0.817
DL	0.993	0.999	0.423	1.965	0.999	0.863
DLO	0.990	0.999	0.275	2.190	0.999	0.769
HS	0.993	1.000	0.323	1.469	1.000	0.823
HSO	0.989	1.000	0.233	1.505	1.000	0.735
RHS	0.994	1.000	0.377	1.720	1.000	0.869
RHSO	0.989	1.000	0.212	1.688	1.000	0.740
Condition: $\rho = 0.95$						
BP	0.969	0.973	0.142	6.425	0.973	0.858
BPO	0.987	0.995	0.092	7.523	0.995	0.788
D2	1.000	1.000	0.031	3.554	1.000	0.994
D2O	0.996	1.000	0.062	5.119	1.000	0.894
DL	0.998	1.000	0.040	3.422	1.000	0.948
DLO	0.994	1.000	0.042	3.935	1.000	0.863
HS	0.985	1.000	0.065	1.363	1.000	0.629
HSO	0.984	1.000	0.056	1.493	1.000	0.606
RHS	0.992	1.000	0.042	1.571	1.000	0.796
RHSO	0.989	1.000	0.050	1.744	1.000	0.725

References

Aguilar, J. E. and Bürkner, P.-C. (2023). "Intuitive joint priors for Bayesian linear multilevel models: The R2D2M2 prior." *Electronic Journal of Statistics*, 17(1): 1711 – 1767.
URL https://doi.org/10.1214/23-EJS2136 2, 7, 12, 13, 14, 28, 33

— (2025). "Generalized Decomposition Priors on R2." *Bayesian Analysis*, 1 – 34.
URL https://doi.org/10.1214/25-BA1524 5, 28

Anderson, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*. Wiley. 29

Armagan, A., Clyde, M., and Dunson, D. B. (2011). "Generalized beta mixtures of Gaussians." In *Advances in neural information processing systems*, 523–531. 5, 13

Bai, R. and Ghosh, M. (2019). "Large-scale multiple hypothesis testing with the normal-beta prime prior." *Statistics*, 53(6): 1210–1233.
URL https://doi.org/10.1080/02331888.2019.1662017 5, 7, 13, 19, 27

— (2021). "On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models." *Statistica Sinica*.
URL http://arxiv.org/abs/1807.06539 27

Banerjee, S., Castillo, I., and Ghosal, S. (2021). "Bayesian inference in high-dimensional models." ArXiv:2101.04491 [math].
URL http://arxiv.org/abs/2101.04491 18

Banerjee, S. and Ghosal, S. (2014). "Posterior convergence rates for estimating large precision matrices using graphical models." *Electronic Journal of Statistics*, 8(2): 2111–2137. 18

Benjamini, Y. and Hochberg, Y. (1995). "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1): 289–300.
URL http://www.jstor.org/stable/2346101 20

Berger, J. (1985). *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer. 20

Bhadra, A., Datta, J., Polson, N. G., and Willard, B. (2016). "Default Bayesian analysis with global-local shrinkage priors." *Biometrika*, 103(4): 955–969.
URL http://www.jstor.org/stable/26363497 2

— (2019). "Lasso Meets Horseshoe: A Survey." *Statistical Science*, 34(3): 405 – 427.
URL https://doi.org/10.1214/19-STS700 3

Bhattacharya, A., Chakraborty, A., and Mallick, B. K. (2016). "Fast sampling with Gaussian scale mixture priors in high-dimensional regression." *Biometrika*, 103(4): 985–991.
URL https://doi.org/10.1093/biomet/asw042 2

Bhattacharya, A., Pati, D., Pillai, N. S., and Dunson, D. B. (2015). "Dirichlet–Laplace Priors for Optimal Shrinkage." *Journal of the American Statistical Association*, 110(512): 1479–1490.
URL https://doi.org/10.1080/01621459.2014.960967 1, 2, 5, 11, 19, 20,

24, 27

Brooks, Gelman, S., and Jones, A. (2011). *Handbook of Markov Chain Monte Carlo*. Chapman and Hall/CRC, 1 edition. 19, 20

Brown, P. J. and Brown, P. J. (1994). *Measurement, Regression, and Calibration*. Oxford Statistical Science Series. Oxford, New York: Oxford University Press. 28

Brown, P. J. and Griffin, J. E. (2010). "Inference with normal-gamma prior distributions in regression problems." *Bayesian Analysis*, 5(1): 171 – 188.
URL https://doi.org/10.1214/10-BA507 1, 5, 7, 11, 20, 28

Bühlmann, P. and Van De Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Series in Statistics. Berlin, Heidelberg: Springer.
URL https://link.springer.com/10.1007/978-3-642-20192-9 2

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). "Stan: A Probabilistic Programming Language." *Journal of Statistical Software*, 76(1): 1–32.
URL https://www.jstatsoft.org/index.php/jss/article/view/v076i01 19

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2009). "Handling Sparsity via the Horseshoe." In van Dyk, D. and Welling, M. (eds.), *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, 73–80. PMLR.
URL https://proceedings.mlr.press/v5/carvalho09a.html 33

— (2010). "The horseshoe estimator for sparse signals." *Biometrika*, 97(2): 465–480.
URL http://www.jstor.org/stable/25734098 1, 2, 5, 7, 11, 12, 13, 19, 20, 24, 27

Casella, G., Ghosh, M., Gill, J., and Kyung, M. (2010). "Penalized regression, standard errors, and Bayesian lassos." *Bayesian Analysis*, 5(2).
URL https://projecteuclid.org/journals/bayesian-analysis/volume-5/issue-2/Penalized-regression-standard-errors-and-Bayesian-lassos/10.1214/10-BA607.full 2, 3

Castillo, I. (2024). "Bayesian nonparametric statistics, St-Flour lecture notes."
URL http://arxiv.org/abs/2402.16422 3

Castillo, I., Schmidt-Hieber, J., and van der Vaart, A. (2015). "Bayesian Linear Regression with Sparse Priors." *The Annals of Statistics*, 43(5): 1986–2018.
URL https://www.jstor.org/stable/43818568 2, 3

Castillo, I. and Vaart, A. v. d. (2012). "Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences." *The Annals of Statistics*, 40(4): 2069 – 2101.
URL https://doi.org/10.1214/12-AOS1029 2, 4, 12

Filzmoser, P. (2023). *chemometrics: Multivariate Statistical Analysis in Chemometrics*.
URL https://CRAN.R-project.org/package=chemometrics 23

Gelman, A. (2006). "Prior distributions for variance parameters in hierarchical

models (comment on article by Browne and Draper)." *Bayesian Analysis*, 1(3): 515 – 534.
URL https://doi.org/10.1214/06-BA117A 2, 27

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., and Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis. 1, 2, 6, 20

Gelman, A., Hill, J., and Vehtari, A. (2020). *Regression and Other Stories*. Analytical Methods for Social Research. Cambridge University Press. 1, 2

George, E. I. and McCulloch, R. E. (1993). "Variable Selection Via Gibbs Sampling." *Journal of the American Statistical Association*, 88(423): 881–889.
URL https://www.jstor.org/stable/2290777 2, 6

— (1995). "Stochastic search variable selection." *Markov Chain Monte Carlo in Practice*. 6

Ghosal, S., Ghosh, J. K., and Vaart, A. W. v. d. (2000). "Convergence rates of posterior distributions." *The Annals of Statistics*, 28(2): 500 – 531.
URL https://doi.org/10.1214/aos/1016218228 3

Ghosh, J. K. (2003). *Bayesian Nonparametrics*. Springer Series in Statistics Ser. New York, NY: Springer New York. 3

Ghosh, P. and Chakrabarti, A. (2017). "Asymptotic Optimality of One-Group Shrinkage Priors in Sparse High-dimensional Problems." *Bayesian Analysis*, 12(4): 1133 – 1161.
URL https://doi.org/10.1214/16-BA1029 2, 4, 7

Giraud, C. (2014). *Introduction to High-Dimensional Statistics*. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis. 1, 2, 16

Goplerud, M. (2021). "Modelling Heterogeneity Using Bayesian Structured Sparsity."
URL http://arxiv.org/abs/2103.15919 3

Griffin, J. E. and Brown, P. J. (2021). "Bayesian global-local shrinkage methods for regularisation in the high dimension linear model." *Chemometrics and Intelligent Laboratory Systems*, 210: 104255.
URL https://www.sciencedirect.com/science/article/pii/S016974392100023X 28

Griffin, M. and Hoff, P. D. (2024). "Structured Shrinkage Priors." *Journal of Computational and Graphical Statistics*, 33(1): 1–14.
URL https://www.tandfonline.com/doi/citedby/10.1080/10618600.2023.2233577 2, 3, 6, 7, 25

Hagar, L. and Stevens, N. T. (2024). "Posterior Ramifications of Prior Dependence Structures."
URL http://arxiv.org/abs/2312.06437 3

Hahn, P. R. and Carvalho, C. M. (2015). "Decoupling Shrinkage and Selection in Bayesian Linear Models: A Posterior Summary Perspective." *Journal of the American Statistical Association*, 110(509): 435–448.
URL https://doi.org/10.1080/01621459.2014.993077 3, 5, 6

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The Elements of Statistical*

Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Science & Business Media. 1

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*. ISSN. CRC Press. 2, 3

Hoerl, A. E. and Kennard, R. W. (1970). "Ridge Regression: Biased Estimation for Nonorthogonal Problems." *Technometrics*, 12(1): 55–67.
URL http://www.jstor.org/stable/1267351 12

Hoffman, M. D. and Gelman, A. (2014). "The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo." *J. Mach. Learn. Res.*, 15(1): 1593–1623. 19

Johnstone, I. M. and Silverman, B. W. (2004). "Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences." *The Annals of Statistics*, 32(4): 1594 – 1649.
URL https://doi.org/10.1214/009053604000000030 2

Lam, C. (2020). "High-dimensional covariance matrix estimation." *Wiley Interdisciplinary Reviews: Computational Statistics*, 12(2).
URL https://onlinelibrary.wiley.com/journal/19390068 16, 17

Ledoit, O. and Wolf, M. (2004a). "Honey, I Shrunk the Sample Covariance Matrix." *The Journal of Portfolio Management*, 30(4): 110–119.
URL https://www.pm-research.com/content/iijpormgmt/30/4/110 17

— (2004b). "A well-conditioned estimator for large-dimensional covariance matrices." *Journal of Multivariate Analysis*, 88(2): 365–411.
URL https://www.sciencedirect.com/science/article/pii/S0047259X03000964 17, 36

Lee, K., Jo, S., and Lee, J. (2022). "The beta-mixture shrinkage prior for sparse covariances with near-minimax posterior convergence rate." *Journal of Multivariate Analysis*, 192: 105067. 18

Lee, K. and Lee, J. (2021). "Estimating large precision matrices via modified Cholesky decomposition." *Statistica Sinica*, 31(1): 173–196. 18

Li, X., Zhao, T., Wang, L., Yuan, X., and Liu, H. (2024). "flare: Family of Lasso Regression."
URL https://cran.r-project.org/web/packages/flare/index.html 24

Li, Y. and Clyde, M. A. (2018). "Mixtures of g-Priors in Generalized Linear Models." *Journal of the American Statistical Association*, 113(524): 1828–1845.
URL https://doi.org/10.1080/01621459.2018.1469992 2, 5, 6

Li, Y., Craig, B. A., and Bhadra, A. (2019). "The graphical horseshoe estimator for inverse covariance matrices." *Journal of Computational and Graphical Statistics*, 28(3): 747–757. 18

Liang, F., , P., Rui, , M., German, , C., Merlise A, and Berger, J. O. (2008). "Mixtures of g Priors for Bayesian Variable Selection." *Journal of the American Statistical Association*, 103(481): 410–423.
URL https://doi.org/10.1198/016214507000001337 2, 6

Maruyama, Y. and George, E. I. (2011). "Fully Bayes factors with a generalized g-prior." *The Annals of Statistics*, 39(5): 2740 – 2765.
URL https://doi.org/10.1214/11-AOS917 2, 5, 6

Marčenko, V. A. and Pastur, L. A. (1967). "Distribution of eigenvalues for some sets of random matrices." *Mathematics of the USSR-Sbornik*, 1(4): 457. 17

Mitchell, T. J. and Beauchamp, J. J. (1988). "Bayesian Variable Selection in Linear Regression." *Journal of the American Statistical Association*, 83(404): 1023–1032.
URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1988.10478694 2

Neal, R. M. (2011). *MCMC using Hamiltonian dynamics*.
URL http://arxiv.org/abs/1206.1901 19

Pas, S. L. v. d. (2021). "Theoretical guarantees for the horseshoe and other global-local shrinkage priors." In *Handbook of Bayesian Variable Selection*, 133–160. Chapman and Hall/CRC. 2, 4

Pas, S. L. v. d., Salomond, J.-B., and Schmidt-Hieber, J. (2016). "Conditions for posterior contraction in the sparse normal means problem." *Electronic Journal of Statistics*, 10(1): 976 – 1000.
URL https://doi.org/10.1214/16-EJS1130 2, 4, 7

Pas, S. L. v. d., Szabó, B., and Vaart, A. v. d. (2017). "Uncertainty Quantification for the Horseshoe (with Discussion)." *Bayesian Analysis*, 12(4): 1221 – 1274.
URL https://doi.org/10.1214/17-BA1065 2, 3, 13

Pati, D., Bhattacharya, A., Pillai, N. S., and Dunson, D. B. (2014). "Posterior contraction in sparse Bayesian factor models for massive covariance matrices." *The Annals of Statistics*, 42(3): 1102–1130. 18

Pauger, D. and Wagner, H. (2019). "Bayesian Effect Fusion for Categorical Predictors." *Bayesian Analysis*, 14(2): 341 – 369.
URL https://doi.org/10.1214/18-BA1096 3

Piironen, J., Paasiniemi, M., and Vehtari, A. (2020). "Projective inference in high-dimensional problems: Prediction and feature selection." *Electronic Journal of Statistics*, 14(1): 2155 – 2197.
URL https://doi.org/10.1214/20-EJS1711 5

Piironen, J. and Vehtari, A. (2017). "Sparsity information and regularization in the horseshoe and other shrinkage priors." *Electronic Journal of Statistics*, 11(2): 5018 – 5051.
URL https://doi.org/10.1214/17-EJS1337SI 1, 2, 12, 13, 14, 19, 24

Robert, C. (2007). *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*. Springer Texts in Statistics. Springer New York. 19

Robert, C. P. and Casella, G. (2004). *Monte Carlo Statistical Methods*. Springer Texts in Statistics. New York, NY: Springer.
URL http://link.springer.com/10.1007/978-1-4757-4145-2 2, 6, 11

Ročková, V. and George, E. I. (2018). "The Spike-and-Slab LASSO." *Journal of the American Statistical Association*, 113(521): 431–444.
URL https://doi.org/10.1080/01621459.2016.1260469 1

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields: Theory and Applications*. New York: Chapman and Hall/CRC. 35

Sagar, K., Banerjee, S., Datta, J., and Bhadra, A. (2024). "Precision matrix

estimation under the horseshoe-like prior–penalty dual." *Electronic Journal of Statistics*, 18(1): 1–46. 18

Scheetz, T. E., Kim, K.-Y. A., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C., and Stone, E. M. (2006). "Regulation of gene expression in the mammalian eye and its relevance to eye disease." *Proceedings of the National Academy of Sciences of the United States of America*, 103(39): 14429–14434.
URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC1636701/ 24

Silverstein, J. W. (1995). "Strong Convergence of the Empirical Distribution of Eigenvalues of Large Dimensional Random Matrices." *Journal of Multivariate Analysis*, 55(2): 331–339.
URL https://www.sciencedirect.com/science/article/pii/S0047259X85710834 17

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). "Penalising Model Component Complexity: A Principled, Practical Approach to Constructing Priors." *Statistical Science*, 32(1): 1 – 28.
URL https://doi.org/10.1214/16-STS576 2, 3

Stan Development Team (2024). "Stan Modeling Language Users Guide and Reference Manual, Version 2.36."
URL http://mc-stan.org/ 19

Tadesse, M. and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press. 2, 6, 13

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." *Journal of the Royal Statistical Society. Series B (Methodological)*, 267–288. 2, 3

Vehtari, A., Gelman, A., and Gabry, J. (2016). "Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC." *Statistics and Computing*, 27(5): 1413–1432.
URL http://dx.doi.org/10.1007/s11222-016-9696-4 19, 24, 25

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). "Rank-Normalization, Folding, and Localization: An Improved $\widehat{R}$ for Assessing Convergence of MCMC (with Discussion)." *Bayesian Analysis*, 16(2): 667 – 718.
URL https://doi.org/10.1214/20-BA1221 20

Vehtari, A. and Ojanen, J. (2012). "A survey of Bayesian predictive methods for model assessment, selection and comparison." *Statistics Surveys*, 6(none): 142 – 228.
URL https://doi.org/10.1214/12-SS102 19

Wainwright, M. J. (2019). *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press. 1

West, M. (1987). "On scale mixtures of normal distributions." *Biometrika*, 74(3): 646–648.
URL https://doi.org/10.1093/biomet/74.3.646 4, 6

Weyl, H. (1912). "Das asymptotische Verteilungsgesetz der Eigenwerte linearer

partieller Differentialgleichungen (mit einer Anwendung auf die Theorie der Hohlraumstrahlung)." *Mathematische Annalen*, 71(4): 441–479.
URL https://doi.org/10.1007/BF01456804 29

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." *Bayesian inference and decision techniques*. 2, 5, 16

— (1996). "Models, prior information, and Bayesian analysis." *Journal of Econometrics*, 75(1): 51–68.
URL https://www.sciencedirect.com/science/article/pii/0304407695017682 2, 16

Zhang, Y. D., Naughton, B. P., Bondell, H. D., and Reich, B. J. (2020). "Bayesian Regression Using a Prior on the Model Fit: The R2-D2 Shrinkage Prior." *Journal of the American Statistical Association*, 0(0): 1–13.
URL https://doi.org/10.1080/01621459.2020.1825449 1, 2, 5, 7, 11, 19, 20, 28, 33

Zhao, P. and Yu, B. (2006). "On Model Selection Consistency of Lasso." *Journal of Machine Learning Research*, 7(90): 2541–2563.
URL http://jmlr.org/papers/v7/zhao06a.html 3

Eidesstattliche Versicherung

Ich erkläre, dass ich die vorliegende Dissertation mit dem Titel

"Intuitive joint priors for high-dimensional Bayesian linear models"

selbstständig verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Alle wörtlich oder inhaltlich übernommenen Stellen habe ich als solche gekennzeichnet.

Ich versichere außerdem, dass ich die beigefügte Dissertation nur in diesem und keinem anderen Promotionsverfahren eingereicht habe und dass diesem Promotionsverfahren keine endgültig gescheiterten Promotionsverfahren vorausgegangen sind. Ferner erkläre ich, dass keine Aberkennung eines bereits erworbenen Doktorgrades vorliegt.

Die Bedeutung der eidesstattlichen Versicherung und die strafrechtlichen Folgen einer unrichtigen oder unvollständigen eidesstattlichen Versicherung sind mir bekannt. Ich versichere an Eides statt, dass ich nach bestem Wissen die reine Wahrheit erklärt und nichts verschwiegen habe.

Ort, Datum

Unterschrift

Ergänzung zur Eidesstattlichen Versicherung (Affidavit) hinsichtlich des Einsatzes von generativen IT-/KI-gestützten Sprachmodellen

Name, Vorname
(surname, first name)

Matrikelnummer
(student ID number)

Titel der Dissertation
(title of the dissertation)

In Ergänzung zur abgegebenen eidesstattlichen Versicherung (Affidavit) versichere ich, dass ich bei der Erstellung der vorliegenden Dissertation betreut von_____
(Erstbetreuer) im Falle des Einsatzes von generativen IT-/KI-gestützten Sprachmodellen diese Werkzeuge in der Übersicht verwendeter Hilfsmittel mit ihrem Produktnamen und der Art des Hilfsmittels, der Art der Verwendung und ggf. den jeweiligen Prompts vollständig aufgeführt habe (siehe beispielhaft nachfolgende Tabelle). Erlaubte Einsatzzwecke sind unter anderem Optimierung selbstverfasster Texte hinsichtlich Grammatik, Rechtschreibung und Schreibstil, Brainstorming bei der Forschungsfrage oder Unterstützung bei der Literaturrecherche, die die Eigenständigkeit bei der Bearbeitung der Abschlussarbeit nicht einschränken. Ferner versichere ich, dass über die genannten Beispiele hinausgehende Verwendung von IT-/KI-gestützten Sprachmodellen mit dem/der Betreuer:in der Arbeit abgesprochen wurde.
Auf die Belehrung hinsichtlich der Folgen von Täuschungsversuchen und der Abgabe einer falschen Versicherung an Eides statt wird auf die eidesstattliche Versicherung und die Promotionsordnung verwiesen. Diese Belehrung habe ich zur Kenntnis genommen.

In addition to the affidavit submitted, I hereby affirm that I have used generative IT-/AI-supported language models in the preparation of this dissertation under the supervision of

_____
(first supervisor) and that I have listed these tools in full in the overview of tools used with their product name and the type of tool, the type of use and, if applicable, the respective prompts (see table below as an example). Permitted uses include optimization of self-authored texts with regard to grammar, spelling and writing style, brainstorming on the research question or support in literature research that does not restrict independence in the processing of the thesis. Furthermore, I assure that the use of IT-/AI-supported language models beyond the examples mentioned has been discussed with the supervisor(s) in the dissertation.
Please refer to the affidavit and the doctoral regulations for information on the consequences of attempted cheating and making a false declaration in lieu of an oath. I have taken note of these instructions.

Verwendete IT-/KI-gestützte Sprachmodelle IT/AI-supported language models used

Produktname Productname	**Art des Hilfsmittels** Type of tool	**Art der Verwendung** Type of use	**Prompt**	**Textstelle(n)** Text passage(s)

Ort, Datum
(place, date)

Unterschrift
(signature)