technische universität dortmund

fakultät statistik

Non-Linear Modeling and Structured Variable Selection in Environmental and Biomedical Data

DISSERTATION

in partial fulfillment of the requirements for the degree of

Doktor der Naturwissenschaften

submitted to the

Department of Statistics

TU Dortmund University

by

Dayasri Ravi

April, 2025

Primary referee: Prof. Dr. Andreas Groll

Secondary referee: JProf. Dr. Christian Staerk

Commission chairperson: Prof. Dr. Jörg Rahnenführer

Assessor: Prof. Dr. Tamara Schikowski

Date of the oral examination: 21.05.2025

Abstract

This thesis addresses three key challenges in the analysis of biological and environmental datasets. First, potential influence factors such as environmental or clinical variables are often modeled as having linear effects on health outcomes. In practice, however, these effects can be non-linear and may involve complex interactions between variables. Despite their relevance, non-linear modeling techniques remain underutilized in the environmental health domain. Second, the problem of variable selection is complicated because relevant predictors are often naturally grouped, such as molecular data, environmental variables, clinical information, family history, and genetic or pathological markers. Traditional variable selection methods often fail to account for this grouping structure, leading to the over-representation of certain groups and the neglect of others, especially low-dimensional but clinically meaningful variables. Third, this limitation is particularly problematic in time-to-event prediction, where clinical and pathological features can substantially impact patient survival outcomes.

This cumulative thesis comprises three studies with five contributed articles that aim to overcome these methodological challenges. The first study investigates the joint effects of ambient temperature and air pollution on systolic and diastolic blood pressure in elderly German women. Using generalized additive models (GAMs), the study captures non-linear exposure-response relationships and complex interactions. The second project provides a new methodological contribution and introduces a novel variant of the Exclusive Lasso for high-dimensional data with grouped variables, such as multi-omics datasets. Smooth approximations are applied to address the non-differentiability of the group-wise L_1-norm, enabling efficient optimization using Newton-based methods. Unlike the conventional Exclusive Lasso, the proposed method does not force selection from every group, allowing for greater sparsity and improved performance. The third study extends this regularization technique to time-to-event data by incorporating Exclusive Lasso into the Cox proportional hazards model. This allows the integration of multiple heterogeneous data types, including gene expression and clinical variables, while preserving the grouping structure. The method is applied to a real-world cancer dataset, showing improved survival prediction and ensuring that low-dimensional but important clinical variables are retained in the model.

Acknowledgements

I would like to express my deep gratitude to my primary supervisor, Prof. Dr. Andreas Groll, for his immense support and constant encouragement throughout my PhD journey. I am deeply thankful for the academic freedom to explore and pursue independent research ideas and for always being available when guidance is needed. His optimism and faith in my work helped me stay focused and motivated through all stages of this journey.

Thanks to Prof. Dr. Tamara Schikowski and her team at the IUF for their interdisciplinary guidance and support. It was an excellent opportunity to collaborate and gain valuable insights across disciplines.

I am immensely grateful to have worked within the Research Training Group 2624: *Statistical Methods for High-Dimensional Data in Toxicology*. I sincerely thank Prof. Dr. Jörg Rahnenführer and the RTG team for creating such a supportive and inspiring research environment.

Moreover, I want to thank JProf. Dr. Christian Staerk for kindly agreeing to serve as the second referee of this thesis. I sincerely appreciate his time and consideration.

Thanks are due to the Technical University of Munich, where this academic journey began in a foreign land—now a place I call home.

I am grateful to the Department of Statistics, University of Madras, and especially Dr. M. Subbiah, for the early foundation in statistics and research that set the stage for all that followed.

I am forever grateful to my family for their unwavering love and support. To my father, Ravi, thank you for always being the first to read my work and for being my pillar of strength. To my sister, Rajyasri, your presence made this journey lighter, warmer, and more joyful. To Samyajoy, with whom I've shared this journey, thank you for your constant support and belief in me and for answering all those crazy statistical doubts at odd hours. I truly couldn't have done this without the three of you.

Finally, to my mother, Nandini—my first mathematics teacher and my guiding star—thank you for being the inspiration behind it all.

List of Contributed Articles

This cumulative thesis is based on the following five manuscripts, which are referred to in the thesis by their respective Roman numerals.

I. **Ravi, D., Groll, A., and Schikowski, T.** (2022). Non-linear modelling of systolic and diastolic blood pressures via environmental factors. *Proceedings of the 36th International Workshop on Statistical Modelling, 1, 294–299.*
Contribution of the author: The author of this thesis conducted the initial data analysis and carried out the statistical analyses. The author primarily wrote the manuscript.

II. **Ravi, D., Groll, A., Wigmann, C., Singh, N., and Schikowski, T.** (2024). Complex synergistic effects of air pollution and temperature on blood pressure: Evidence from the SALIA cohort study. *Submitted (under review).*
Contribution of the author: The author of this thesis contributed to the conceptualization of the study, performed statistical analysis, and conducted an extensive review of the relevant literature. The first draft of the manuscript was written independently by the author, with critical input and refinements from the co-authors.

III. **Ravi, D., and Groll, A.** (2024). Optimizing variable selection in multi-omics datasets: A focus on exclusive lasso. *International Workshop on Statistical Modelling.* Cham: Springer Nature Switzerland, DOI: `https://doi.org/10.1007/978-3-031-65723-8_22`
Contribution of the author: The author of this thesis developed the methodological approach, performed the data simulations, and wrote the majority of the manuscript, with revisions and suggestions provided by the co-author.

IV. **Ravi, D., and Groll, A.** (2025). A Newton-based variant of Exclusive Lasso for improved sparse solutions. *Comput Stat.*, DOI: `https://doi.org/10.1007/s00180-025-01630-5`
Contribution of the author: The author of this thesis developed the core concept for the proposed model, implemented the software package, conducted the simulation study and data analysis, and drafted the initial version of the manuscript. Comments from insightful discussions with the co-author were incorporated.

V. **Ravi, D., and Groll, A.** (2025). Time-to-event prediction for grouped variables using Exclusive Lasso. *arXiv preprint*: arXiv: 2504.01520
Contribution of the author: The author of this thesis developed and implemented a new

methodological approach, carried out data simulations, analyzed real-world data, and wrote the majority of the manuscript. Writing and revisions were carried out collaboratively with the co-author.

Contents

Part I

Summary of Thesis Work

1. Introduction

The Generalized Linear Model (GLM) is one of the most widely used statistical approaches for examining the relationship between environmental factors and epidemiological or clinical outcomes. In GLMs, a linear combination of environmental predictors is linked to clinical outcomes through a specified link function, enabling the use of various probability distributions (e.g., normal, binomial, Poisson) to accommodate non-normal error structures and better capture complex health-environment interactions. However, assuming a strictly linear relationship between predictors and outcomes can be overly restrictive, as environmental effects are often more complex and non-linear, especially at higher exposure levels. Since the precise nature of these non-linear effects is typically unknown, allowing the data to determine the functional form automatically within a statistical framework is preferable. Generalized Additive Models (GAMs) address this limitation by extending GLMs to model predictor-response relationships non-linearly using smooth functions (see, for example, Wood, 2017; Eilers and Marx, 2021). Instead of imposing a predefined parametric structure, GAMs adopt a semi-parametric approach, allowing the data to shape the relationship between predictors and outcomes dynamically.

In recent years, GAMs have become increasingly popular in environmental health research. Ravindra et al. (2019) conducted a comprehensive review of the application of GAMs for studying air pollution and its health effects, highlighting both their advantages and limitations. The independent effects of environmental factors such as temperature and air pollution on human health have been widely examined using GAMs (Pearce et al., 2011; Li et al., 2018, 2019). However, there is growing recognition of the need to move beyond single-pollutant models towards multi-pollutant frameworks (Dominici et al., 2010; Mauderly et al., 2010). Multiple pollutants can be simultaneously included in a GAM, allowing the effect of one pollutant to be estimated while adjusting for the presence of others in the model.

A significant challenge in this setting is the high correlation between many air pollutants arising from shared temporal and spatial emission patterns and the influence of common meteorological conditions. For instance, Chen et al. (2008) and Kan et al. (2010) found that the estimated effect of particulate matter (PM_{10}) on daily respiratory mortality became negative when nitrogen dioxide (NO_2) was included in the model. This emphasizes the need to account for correlations between pollutants in multi-pollutant models, as ignoring them may lead to collinearity issues, inflated variance, or biased estimates.

Another important challenge in environmental modeling is capturing potential interaction effects between environmental variables. Evidence indicates that the combined impact of multiple exposures may not be simply additive but may differ from the sum of individual effects estimated in single-pollutant models (Mauderly and Samet, 2009). A common approach to address this is through multivariate models that include both main effects for each pollutant and pairwise interaction terms (Dominici et al., 2010). However, it is well known that statistical power to detect two-way interactions is often limited, especially in the absence of strong interaction effects. Accounting for higher-order interactions presents an even more significant challenge. Moreover, identifying all possible pairwise interactions becomes increasingly difficult as the number of pollutants grows. An alternative strategy involves stratifying the data based on one environmental variable and including a multiplicative interaction term to assess effect modification (Cheng and Kan, 2012). This may, however, lead to a loss of power that can arise from stratification (Royston et al., 2006).

To address these challenges, Sun et al. (2013) explored statistical methods designed to accommodate multiple pollutants. They proposed a two-step procedure: First, pollutant selection is carried out using classification and regression trees (CART;Breiman et al., 2017), followed by dimension reduction through algorithms such as the Least Absolute Shrinkage and Selection Operator (Lasso; Tibshirani, 1996). However, this approach does not capture non-linear main effects or complex interaction structures among pollutants. Therefore, there remains a clear need for modeling strategies capable of flexibly capturing both non-linear exposure-response relationships and potential interactions among environmental variables. The tensor product smooths address this need by extending univariate smoothing techniques to accommodate multiple covariates, enabling the estimation of complex, multidimensional relationships between exposures and outcomes.

To this end, Wood (2006) developed a general framework for constructing low-rank tensor product smooths, making them computationally efficient and applicable in the GAM framework. In environmental health modeling, tensor product smooths are particularly advantageous as they allow for the simultaneous inclusion of several correlated pollutants while capturing potential synergistic effects and mitigating issues related to multicollinearity.

This idea is explored in Articles I and II of this thesis, which investigate the potential synergistic effects of mean temperature and air pollutants, specifically particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), and ozone (O_3), on systolic and diastolic blood pressure in elderly German women. A GAM incorporating bivariate tensor product splines was employed to capture the complex joint effects of temperature and air pollutants. Blood pressure was modeled as a smooth bivariate function of temperature and pollutant concentrations, with adjustment for common confounders such as age, smoking status, and living conditions. These articles highlight the importance of non-linear modeling approaches for environmental variables in relation to blood pressure outcomes and emphasize the critical role of GAMs in revealing and interpreting complex interaction effects among environmental exposures.

In our analysis, blood pressure was primarily explained using environmental factors. However, several other influential variables play an important role in understanding human health beyond environmental exposures. For example, genetic predispositions, clinical and pathological conditions, psychosocial stressors, and dietary habits may also affect blood pressure. This presents an additional

challenge of selecting the most relevant variables from a large and diverse set of potential predictors. An important consideration is that these variables are often naturally grouped based on their origin or type, such as genetic, clinical, environmental, or behavioral. This requires modeling approaches that can effectively incorporate this structured grouping.

A common approach for variable selection that accounts for the grouping structure among predictors is the Group Lasso method (Yuan and Lin, 2006). Group Lasso applies an $L_{2,1}$-norm penalty, which combines the L_1-norm across groups to induce sparsity at the group level and the L_2-norm within each group to jointly regularize variables belonging to the same group. This encourages entire groups of variables to be either selected or excluded from the model. Despite its advantages, Group Lasso has notable limitations, particularly in settings where groups of variables are highly correlated, as is often the case in multi-omics datasets. In such situations, the method tends to select variables from only a few dominant groups while often overlooking smaller or lower-dimensional groups, such as clinical variables, which may carry important predictive information. This limitation highlights the need for alternative methods that promote variable selection across all groups, ensuring that relevant variables from each domain are considered regardless of group size or correlation structure.

The Exclusive Lasso offers an alternative to address this limitation by incorporating intra-group sparsity through an $L_{1,2}$-norm penalty. This method promotes sparsity within groups using the L_1-norm, while the L_2-norm across groups relaxes sparsity between them. As a result, the method encourages the selection of at least one variable from each group (Campbell and Allen, 2017). Its applications are growing across various domains, including multi-task feature learning (Zhou et al., 2010), image processing (Zhang et al., 2015), and clustering (Yamada et al., 2017).

Solving the Exclusive Lasso problem is computationally more demanding due to its composite penalty, which involves the L_1-norm applied within groups. To address this, several optimization strategies have been developed. These include coordinate descent and proximal gradient methods employing soft-thresholding techniques (Campbell and Allen, 2017), as well as proximal point algorithms based on dual Newton methods (Lin et al., 2020). Other strategies use iterative re-weighted schemes (Kong et al., 2014; Sun et al., 2020). Another line of work reformulates the problem into a standard Lasso setting and applies a bisection algorithm that takes advantage of the piecewise linearity of the Lasso solution path (Sun et al., 2020). More recently, a variant based on the fast iterative shrinkage-thresholding algorithm (FISTA) has also been proposed to improve computational efficiency (Huang and Liu, 2018).

In Articles III and IV of this thesis, an alternative estimation algorithm to Exclusive Lasso is established. A novel strategy is introduced to address the $L_{1,2}$-norm-based penalty by employing a smooth approximation, whereby the L_1-norm at the group level is reformulated into a differentiable expression. The proposed approach is designed to be compatible with a wide range of loss functions, thereby enabling its application across diverse modeling frameworks. Once the penalty term is smoothed into a continuous and differentiable form, optimization is carried out efficiently using the Newton-based method. In this work, two specific smoothing functions—a quadratic and a sigmoid approximation—are adopted to approximate the $L_{1,2}$-norm. Related efforts have previously explored smoothing techniques for the L_1-norm using Newton-based optimization algorithms (Schmidt et al., 2007; Nkansah et al., 2021).

A major limitation of the standard Exclusive Lasso is its enforcement of selecting at least one variable per group, even when it contains no informative predictors. In Article IV, this issue is addressed by introducing a more flexible sparsity structure in the proposed approach. Rather than requiring selection within every group, sparsity is controlled through a penalty parameter, and a small threshold is applied to eliminate variables from non-informative groups. This allows uninformative groups to be excluded more effectively. As a result, the optimization process is simplified while maintaining sparsity both within and between groups. In addition, combining thresholding with appropriate initialization leads to more stable coefficient paths compared to the conventional Exclusive Lasso. The proposed method is applied to real-world datasets from finance and biology. In both applications, the advantages of using the Exclusive Lasso over the Group Lasso are demonstrated.

The final part of this thesis focuses on extending the proposed methodology to time-to-event modeling. This extension is critical as many environmental and biological datasets involve survival outcomes. A key question is how to incorporate the grouping structure of variables into survival models. For instance, multi-omics datasets often include variables from distinct domains, such as genomics, transcriptomics, and proteomics. Advances in high-throughput technologies have led to the availability of high-dimensional molecular data, which are frequently accompanied by phenotypic data, including traditional clinical and environmental variables.

Effectively integrating these diverse data sources into survival analysis is challenging. Two common strategies are used in practice. The first is a naive approach, where all variables are combined into a single block, disregarding their source. A Cox proportional hazards model (Cox, 1972) is then fitted, and variable selection is typically performed using the Lasso (Tibshirani, 1997). However, the approach does not account for the group structure, which may result in underrepresenting or omitting lower-dimensional but clinically relevant groups. Such exclusion can reduce prediction accuracy, as several studies suggest that combining clinical and omics variables often yields better performance than using either type alone (Binder and Schumacher, 2008; Bøvelstad et al., 2009; Herrmann et al., 2021). The second strategy involves separate analysis, in which each variable group is analyzed independently before merging the results. A more recent method in this category is the Integrative L_1-Penalized Regression with Penalty Factors (IPF-Lasso; Boulesteix et al., 2017), which allows for group-specific penalties based on prior knowledge or data-driven procedures. Despite these advancements, separate analysis does not account for potential dependencies between groups, and most existing methods do not ensure variable selection from all groups. Consequently, vital information may be excluded from the final model.

In the final manuscript, the Exclusive Lasso regularization is extended to the Cox Proportional Hazards (PH) model. A real-world gene expression dataset is used for survival prediction in bladder cancer. The proposed Exclusive Lasso method is compared to several state-of-the-art statistical approaches that incorporate grouping structures within the Cox model framework. The variable selection results are analyzed across different groups in order to evaluate the effectiveness of each method. The Exclusive Lasso demonstrates superior performance, as informative features are consistently selected from all variable groups, ensuring that relevant biological signals are comprehensively represented.

The structure of this thesis is organized as follows. Chapter 2 provides the necessary statistical and

methodological background, offering a general overview of the techniques that are either applied or further developed throughout the thesis. Chapter 3 presents a summary of the individual research articles, emphasizing the key contributions and novel aspects of each work. Finally, Chapter 4 presents a concluding discussion and outlook on future research, followed by the full versions of the articles.

2. Statistical Methods

This chapter provides a general methodological foundation for the methods utilized or further developed in this thesis.

2.1 Generalized Additive Models

Generalized Additive Models (GAMs; Hastie and Tibshirani, 1986) are a class of semi-parametric models that generalize the structure of Generalized Linear Models (GLMs) by allowing the linear predictor to be an additive combination of smooth functions of the covariates.

Let x_{ij} denote the observed value of the j-th covariate for the i-th observation, where $i = 1, \ldots, n$ and $j = 1, \ldots, p$. Let y_i be an observation on the random variable Y_i, which follows an exponential family distribution. The conditional expectation $\mu_i = \mathbb{E}(Y_i \mid \mathbf{x}_i)$ is linked to an additive predictor through a known link function $g : \mathbb{R} \to \mathbb{R}$, such that

$$g(\mu_i) = \mathbf{A}_i\boldsymbol{\gamma} + f_1(x_{i1}) + f_2(x_{i2}) + \ldots, \quad \text{for } i = 1, \ldots, n,$$

where $\mathbf{A}_i$ is a known row vector of covariates associated with any strictly parametric component of the model, $\boldsymbol{\gamma}$ is the corresponding parameter vector, and each $f_j : \mathbb{R} \to \mathbb{R}$ is an unknown non-linear, but typically smooth function to be estimated from the data. These functions allow for flexible, potentially non-linear effects of the individual covariates on the response. For the sake of simplicity, we consider a univariate model involving a single covariate $x_i \in \mathbb{R}$ and one smooth function:

$$y_i = f(x_i) + \epsilon_i, \tag{2.1}$$

where $f(\cdot)$ is an unknown smooth function, and ϵ_i are independent random errors for which typically $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$ is assumed.

To make it easier to estimate the function $f(\cdot)$, we write it in a form that turns Equation (2.1) into a linear model. This is done by approximating $f(\cdot)$ using a set of known functions called *basis functions*. If $b_q(x)$ is the q-th basis function, we can write:

$$f(x) = \sum_{q=1}^{d} \alpha_q b_q(x), \tag{2.2}$$

where α_q are unknown coefficients. Substituting Equation (2.2) into Equation (2.1) gives a model that is linear in the parameters α_q, which makes it easier to estimate using standard methods.

2.1.1 B-splines

B-splines (Eilers and Marx, 1996, 2021) are a widely used class of basis functions for constructing smooth curves in additive models. They consist of piecewise polynomials defined over a sequence of knots, providing a numerically stable and flexible representation of smooth functions.

Let $l \geq 0$ be the degree of the spline, and suppose the domain $[a, b]$ of x is divided into $m-1$ intervals by an ordered sequence of knots:

$$a = \kappa_1 < \kappa_2 < \cdots < \kappa_m = b.$$

B-spline basis functions are constructed using piecewise polynomials that join smoothly at the knots to maintain a specified level of continuity. Each B-spline basis function consists of $(l+1)$ polynomial segments of degree l, joined in a way that ensures $(l-1)$-times continuous differentiability.

The B-spline basis functions are defined recursively using the Cox–de Boor recursion formula (De Boor, 1972). For degree $l = 0$ (piecewise constant functions), the basis functions are given by:

$$b_{q,0}(x) = \begin{cases} 1, & \text{if } \kappa_q \leq x < \kappa_{q+1}, \\ 0, & \text{otherwise.} \end{cases}$$

For $l \geq 1$, the degree-l basis functions are defined recursively as:

$$b_{q,l}(x) = \frac{x - \kappa_{q-l}}{\kappa_q - \kappa_{q-l}} b_{q-1,l-1}(x) + \frac{\kappa_{q+1} - x}{\kappa_{q+1} - \kappa_{q+1-l}} b_{q,l-1}(x).$$

Using these B-spline basis functions, the function $f(x)$ can be represented as a linear combination of $d = l + m - 1$ basis functions.

B-splines have several properties that make them particularly useful in statistical modeling. Each basis function $b_q(x)$ is nonzero only on a small number of intervals, specifically those determined by $l + 2$ consecutive knots, ensuring local support. This locality property contributes to numerical stability and computational efficiency, as each function is influenced only by a limited range of data points. The structure of B-splines also prevents issues of multicollinearity that commonly arise in polynomial regression. Moreover, their flexibility allows smoothness to be controlled by adjusting the spline degree l and the placement of knots, making them well-suited for nonparametric regression and smoothing applications.

2.1.2 P-splines

As previously noted, the smoothness and the number of B-splines depend on the number of knots. To mitigate this dependence, a roughness penalty is introduced, which discourages excessive flexibility. In particular, rather than relying solely on the standard least squares criterion, we consider a

penalized least squares (PLS) approach that incorporates a penalty on the k-th order differences of the coefficients α_q from Equation (2.2). This leads to the penalized spline framework, or P-splines (Eilers and Marx, 2021), which combines B-splines with a discrete difference penalty.

The focus here is on extending the linear model using B-spline smoothing for Gaussian responses rather than considering the more general framework of GAMs. Given $d = l + m - 1$ B-spline basis functions, the t-th order P-spline estimate minimizes the following penalized residual sum of squares:

$$\text{PLS}(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{q=1}^{d} \alpha_q b_q(x_i) \right)^2 + \lambda \sum_{q=t+1}^{d} (\Delta^t \alpha_q)^2,$$

where $\Delta^t \alpha_q$ denotes the t-th order finite difference of the coefficients α_q, and $\lambda > 0$ is the *smoothing parameter* that governs the trade-off between goodness-of-fit and smoothness.

Let $\mathbf{B} \in \mathbb{R}^{n \times d}$ be the B-spline basis matrix with entries $B_{iq} = b_q(x_i)$, and let $\mathbf{D}^{(t)} \in \mathbb{R}^{(d-t) \times d}$ be the t-th order difference matrix. The penalty can then be written as $\boldsymbol{\alpha}^\top \mathbf{S} \boldsymbol{\alpha}$, where $\mathbf{S} = \mathbf{D}^\top \mathbf{D}$. The penalized objective becomes

$$\text{PLS}(\lambda) = \|\mathbf{y} - \mathbf{B}\boldsymbol{\alpha}\|^2 + \lambda \, \boldsymbol{\alpha}^\top \mathbf{S} \boldsymbol{\alpha}.$$

The minimizer is obtained by solving the penalized normal equations:

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{S} \right)^{-1} \mathbf{B}^\top \mathbf{y}.$$

To estimate the optimal smoothing parameter λ, two commonly used approaches are *Generalized Cross-Validation* (GCV) and *Restricted Maximum Likelihood* (REML).

In GCV, the smoothing parameter is selected by minimizing the criterion:

$$\text{GCV}(\lambda) = \frac{n \|\mathbf{y} - \hat{\mathbf{y}}\|^2}{\left(1 - \frac{1}{n} \text{tr}(\mathbf{H}_\lambda)\right)^2},$$

where $\hat{\mathbf{y}} = \mathbf{H}_\lambda \mathbf{y}$ and $\mathbf{H}_\lambda = \mathbf{B}(\mathbf{B}^\top \mathbf{B} + \lambda \mathbf{S})^{-1} \mathbf{B}^\top$ is the smoothing matrix. The GCV score balances the fit to the data with the complexity of the smoother. To choose the optimal smoothing parameter, GCV is evaluated over a grid of candidate values $\lambda_1, \lambda_2, \ldots, \lambda_L$, and the value that minimizes the GCV score is selected:

$$\hat{\lambda}_{\text{GCV}} = \arg\min_{\lambda} \text{GCV}(\lambda).$$

In contrast to classical cross-validation approaches, GCV has the advantage that it does not require splitting the data into separate training and test sets. This not only simplifies the implementation but also reduces computational cost, as the entire dataset can be used directly in the evaluation process.

REML provides an alternative by interpreting the spline coefficients as random effects and estimating λ by maximizing the restricted likelihood of the marginal distribution of $\mathbf{y}$. This approach generally yields more stable and less biased estimates, particularly in small samples.

For more details on these methods, including their theoretical properties, see Wood (2017).

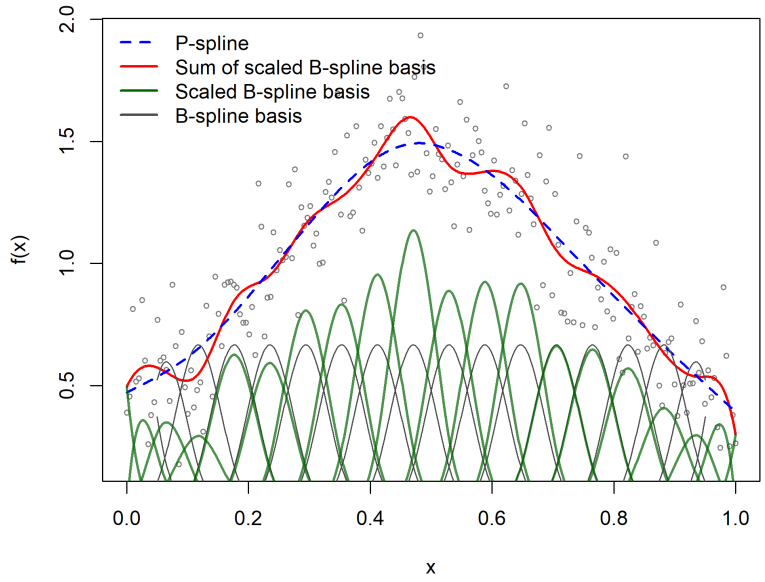

Figure 2.1: Comparison of B-spline (red) and P-spline (blue dashed) fits for noisy observations (black dots) generated from a function. Both methods use a cubic spline basis with 20 degrees of freedom. The green curves represent the scaled B-spline basis functions weighted by their estimated spline coefficients, while the grey curves depict the unscaled B-spline basis functions.

The strength of P-splines lies in their flexibility and computational simplicity. Using a relatively large number of equally spaced knots ensures that the basis system is rich enough to capture the underlying pattern, while the difference penalty avoids overfitting. Moreover, since the penalty is imposed on the coefficients rather than on the function itself, P-splines avoid complex integral calculations associated with traditional smoothing splines. This makes them especially attractive for large datasets and for incorporating smooth terms into GAMs.

Figure 2.1 illustrates how B-splines and P-splines perform smoothing on noisy data. The grey curves correspond to the original, unscaled degree-3 B-spline basis functions. The green curves represent these basis functions scaled by their estimated coefficients, which are linearly combined to produce the final fitted B-spline curve shown in red. This B-spline fit captures more fine-scale variability and tends to overfit the noise due to its flexibility. In contrast, the P-spline fit (blue dashed line) incorporates a smoothness penalty, resulting in a more stable and smoother approximation of the underlying signal.

2.1.3 Tensor Product Smooths

Consider the response variable to be described in terms of a two-dimensional smooth surface $f(x_1, x_2)$, where x_1 and x_2 are continuous covariates. In order to flexibly model the interaction between these

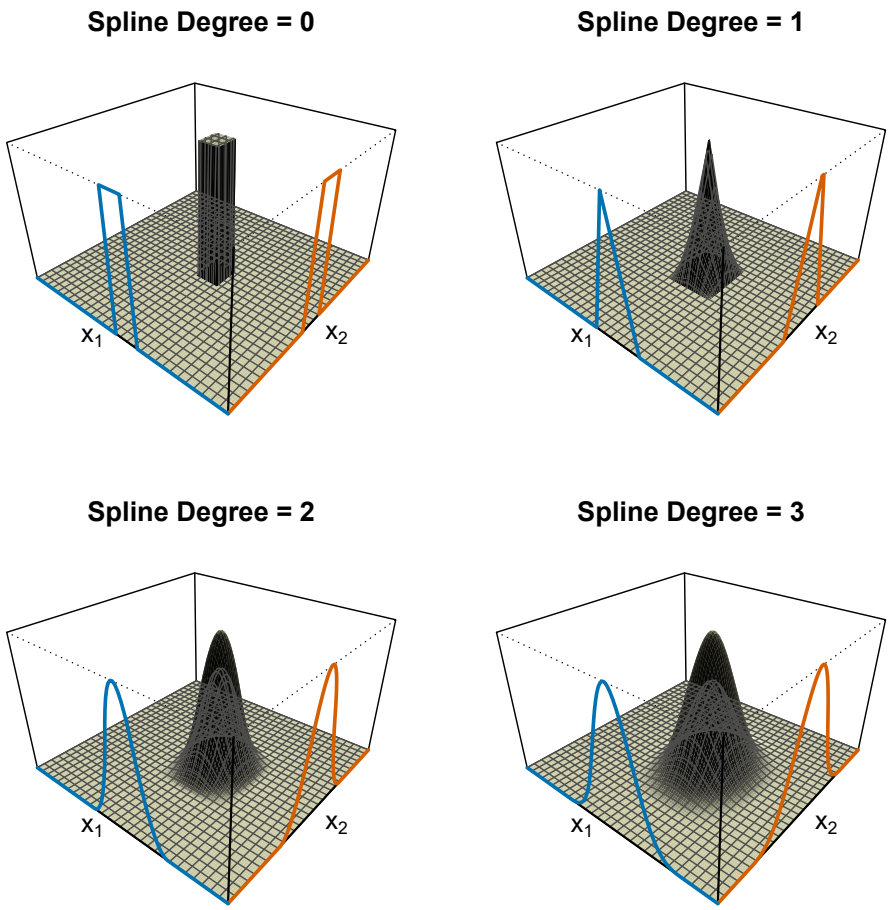

Figure 2.2: Graphical visualization of tensor-product B-spline basis functions constructed from univariate B-splines of degrees 0, 1, 2, and 3, respectively. Each surface represents the product of two selected univariate basis functions defined along the x_1 and x_2 axes. The contributing marginal functions are highlighted as blue and red curves along their respective axes.

covariates, we employ a *tensor product basis* approach. Specifically, we construct univariate basis functions $\{a_q(x_1)\}_{q=1}^{d_1}$ for x_1, and $\{b_r(x_2)\}_{r=1}^{d_2}$ for x_2, such as B-spline or P-spline bases. The resulting bivariate smooth function $f_{12}(x_1, x_2)$ is then expressed as a linear combination of tensor product basis functions:

$$f_{12}(x_1, x_2) = \sum_{q=1}^{d_1} \sum_{r=1}^{d_2} \gamma_{qr}\, a_q(x_1)\, b_r(x_2) \ ,$$

where γ_{qr} are the tensor product coefficients to be estimated from the data. To avoid overfitting and to control the smoothness of the surface in each marginal direction, we apply *double penalization* to the coefficient matrix $\{\gamma_{qr}\}$, with separate roughness penalties in the x_1- and x_2-directions.

Figure 2.2 shows how tensor-product B-spline surfaces are constructed using marginal B-spline basis functions of different degrees. Each surface results from combining one basis function along x_1

with another along x_2, capturing their joint influence. Higher degrees yield smoother, more flexible surfaces. The construction of the tensor product is particularly advantageous when the marginal covariates are measured on different scales or with differing smoothness along each axis.

2.2 Regularization Techniques

GLMs and their flexible extension, GAMs, provide a powerful framework for modeling relationships between predictors and a response variable. However, as the number of predictors increases, especially in high-dimensional settings, these models become prone to overfitting. Moreover, identifying the most relevant variables becomes increasingly challenging. Regularization techniques address these problems by adding constraints to the model, helping to prevent overfitting and making variable selection more effective.

The estimation of the true parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ in a linear model is typically carried out by maximizing the log-likelihood function $\ell(\boldsymbol{\beta})$. Although the intercept term is usually included as the first column of the design matrix and is not subject to penalization, we simplify the presentation by assuming that the response variable has been centered. Consequently, the intercept is omitted from the model without loss of generality.

In regularization frameworks, parameter estimation is modified by introducing a penalty term $P(\boldsymbol{\beta})$, which limits the magnitude of coefficients, encouraging sparser models that are less likely to overfit the data. The resulting *penalized log-likelihood* is given by

$$\ell_{\text{pen}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta}) \, ,$$

where $\lambda \geq 0$ is a regularization parameter that controls the strength of the penalty.

Two widely used choices for the penalty term $P(\boldsymbol{\beta})$ are based on the L_2-norm and the L_1-norm of the coefficient vector.

The L_2-norm penalty is given by

$$P(\boldsymbol{\beta}) = \sum_{k=1}^{p} \beta_k^2 \, ,$$

which leads to *Ridge regression* (Hoerl and Kennard, 1970). The corresponding penalized estimation problem and its corresponding estimates are

$$\hat{\boldsymbol{\beta}}^{\text{ridge}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \ell(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{p} \beta_k^2 \right\} .$$

The L_2-norm penalty shrinks all coefficients toward zero, which helps to stabilize estimation in the presence of multicollinearity and reduce model variance. However, Ridge regression does not perform variable selection, as it does not set any coefficients exactly to zero.

Alternatively, the L_1-norm penalty is defined as

$$P(\boldsymbol{\beta}) = \sum_{k=1}^{p} |\beta_k| \,,$$

and gives rise to the *Least Absolute Shrinkage and Selection Operator* (Lasso; Tibshirani, 1997). The penalized estimation problem and solution under this penalty becomes

$$\hat{\boldsymbol{\beta}}^{\text{lasso}} = \arg\max_{\boldsymbol{\beta} \in \mathbb{R}^p} \left\{ \ell(\boldsymbol{\beta}) - \lambda \sum_{k=1}^{p} |\beta_k| \right\}.$$

The L_1-norm penalty limits the magnitude of the coefficients, encouraging sparser models that are less likely to overfit the data. Unlike Ridge regression, Lasso can shrink some coefficients exactly to zero, thereby performing automatic variable selection in addition to regularization. However, the L_1-norm is not differentiable at zero, which makes a closed-form solution intractable. As a result, numerical methods are required to obtain the solution. In practice, approaches like coordinate descent are commonly used, where all coefficients are fixed except one, and the resulting univariate sub-problem is solved using the soft-thresholding operator.

Figure 2.3 illustrates constraint geometries and residual error contours under various regularization techniques in a two-parameter linear regression setting. The blue ellipses represent contours of the residual sum of squares (RSS), centered at the least squares solution $\hat{\boldsymbol{\beta}}$. The regularized solution corresponds to the point where the RSS contour is tangent to the constraint region. In Ridge regression, the constraint region is a circle, which tends to shrink coefficients toward zero without setting them exactly to zero. In contrast, the Lasso constraint forms a diamond, whose sharp corners increase the likelihood that the RSS contour intersects the constraint boundary at an axis, thereby setting one of the coefficients exactly to zero and promoting sparsity.

Although the Lasso has proven to be highly effective in high-dimensional settings, one of its main limitations is its inability to accommodate grouped structures among covariates naturally. In many biological and epidemiological studies, predictors are often organized into groups based on their origin, function, or measurement domain. In such contexts, performing variable selection at the group level is often desirable rather than on individual predictors.

In this thesis, we focus on variable selection in scenarios where the indices of the true parameter vector $\boldsymbol{\beta}$ are partitioned into non-overlapping, predefined groups. Let $\mathcal{G} = \{1, \ldots, G\}$ denote a collection of such groups, where each group $g \in \mathcal{G}$ corresponds to a subset of indices from $\{1, \ldots, p\}$. The groups are assumed to be disjoint, and their union covers the entire index set:

$$\bigcup_{g \in \mathcal{G}} g = \{1, 2, \ldots, p\} \,.$$

Several extensions of Lasso have been proposed to incorporate this group structure into the regularization framework. In the following, some widely used methods designed for group-level variable selection are discussed.

Elastic Net

The Elastic Net (Zou and Hastie, 2005) is a regularization technique that blends the properties of both Lasso and Ridge regression by combining the L_1- and L_2-norm penalties. The L_1 component encourages sparsity, effectively performing variable selection, while the L_2 component stabilizes the model by encouraging the inclusion of correlated predictors. The Elastic Net penalty function is given by:

$$P(\boldsymbol{\beta}) = \alpha \sum_{k=1}^{p} |\beta_k| + \frac{1}{2}(1-\alpha) \sum_{k=1}^{p} \beta_k^2 \ ,$$

where $\alpha \in [0, 1]$ controls the balance between the two penalties.

A widely used implementation of the Elastic Net is available in the `glmnet` package in R (Simon et al., 2011). In `glmnet`, setting $\alpha = 1$ corresponds to the Lasso, while $\alpha = 0$ yields Ridge regression. A value of $\alpha = 0.5$ results in an equal blend of the two penalties. The regularization parameter λ and the mixing parameter α are typically chosen via cross-validation (CV).

Group Lasso

The Group Lasso (Yuan and Lin, 2006) is designed for situations where predictors are naturally grouped, such as biological or epidemiological data with multi-omics structures. Rather than selecting individual coefficients, it encourages the selection or exclusion of entire groups of variables. The penalty term is defined as:

$$P(\boldsymbol{\beta}) = \sum_{g \in \mathcal{G}} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 \ ,$$

where $\boldsymbol{\beta}_g$ is the subvector of coefficients for group g, and p_g is the size of group g. The square root scaling compensates for unequal group sizes, ensuring fair penalization across groups.

In Figure 2.3, the Group Lasso constraint applies an L_2-norm over the joint group of coefficients. Since both coefficients belong to the same group, the penalty tends to either retain or eliminate them together, exhibiting an "all-in or all-out" behavior.

A suitable implementation is provided exemplarily in the `grpreg` package in R (Breheny and Huang, 2015), which supports grouped penalties for linear and generalized linear models. By default, `grpreg` weights the group penalty by $\sqrt{p_g}$, ensuring comparability across differently sized groups. Again, CV can be used to select the optimal regularization parameter λ.

Sparse Group Lasso

The Sparse Group Lasso (SGL; Simon et al., 2013) extends the Group Lasso by simultaneously encouraging sparsity at both the group and individual levels. This is achieved by combining the Group Lasso penalty with an additional Lasso penalty:

$$P(\boldsymbol{\beta}) = (1-\alpha) \sum_{g \in \mathcal{G}} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 + \alpha \sum_{k=1}^{p} |\beta_k| \ ,$$

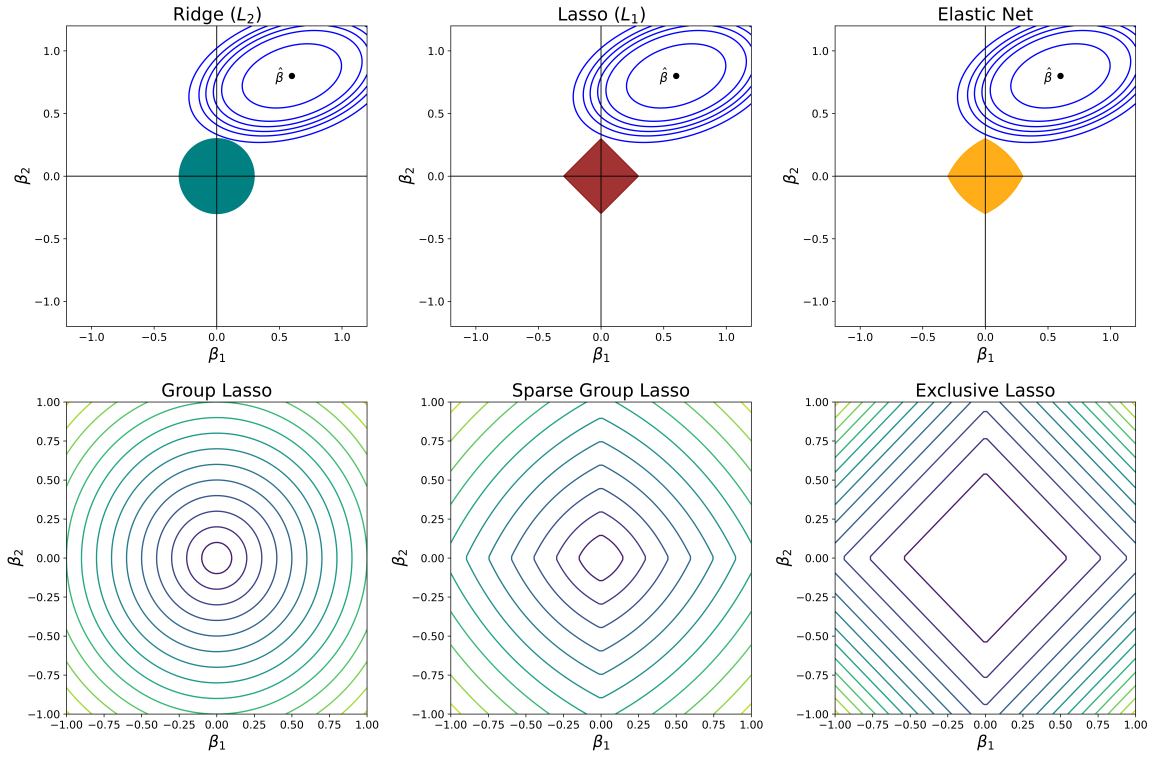

Figure 2.3: Geometrical representation of regularization in a two-parameter linear regression setting. *Top from left to right*: constraint regions for Ridge (L_2), Lasso (L_1), and Elastic Net, alongside the error contours (blue ellipses). *Bottom from left to right*: contours of penalty terms for the Group Lasso, Sparse Group Lasso, and Exclusive Lasso penalties, illustrating their structured sparsity-inducing behavior. Both coefficients are assumed to belong to the same group, which is relevant for the group-based penalties shown in the bottom row.

where $\alpha \in [0, 1]$ governs the trade-off between group-level selection and within-group sparsity.

The `SGL` package in R provides an efficient implementation for both GLM settings and time-to-event predictions (Simon et al., 2019). The default setting in the package uses $\alpha = 0.95$, emphasizing strong within-group sparsity while still preserving group-level structure. This is particularly useful in high-dimensional applications where groups are large and only a few features within each group are expected to be relevant.

Exclusive Lasso

The Exclusive Lasso (Campbell and Allen, 2017) is designed to induce structured sparsity by ensuring that at least one variable is selected from each predefined group. It employs an L_1-penalty within each group to encourage sparsity and an L_2-type structure across groups to avoid entirely discarding

any group. The penalty function is given by:

$$P(\boldsymbol{\beta}) = \frac{1}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} |\beta_k| \right)^2 . \tag{2.3}$$

This formulation encourages selecting a small number of variables within each group while ensuring that all groups are represented in the model. As shown in Figure 2.3, when both variables are part of the same group, the Exclusive Lasso exhibits behavior similar to the Lasso. It lacks a closed-form solution and promotes competition within the group by discouraging the simultaneous selection of multiple variables.

The `ExclusiveLasso` package in `R` provides an implementation of the Exclusive Lasso (Weylandt et al., 2018). Estimation is performed via a coordinate descent algorithm that incorporates a soft-thresholding function to handle the L_1-penalty within each group. An alternative optimization strategy based on proximal gradient descent is also available. A notable drawback of the standard Exclusive Lasso is that it tends to select at least one variable from each group, even when the group contains no informative features. To address this limitation, a modified version of the Exclusive Lasso is proposed in Article IV of this thesis. This is then further extended to the case of time-to-event modeling in Article V.

Figure 2.4 illustrates the distinct sparsity structures induced by various regularization methods when variables are grouped into non-overlapping subsets. Each method encourages a different pattern of coefficient selection, as visualized by the grey (selected) and white (excluded) cells in the figure. The Elastic Net applies element-wise sparsity, treating each variable independently and often resulting in scattered non-zero coefficients. Group Lasso, on the other hand, enforces group-level sparsity by selecting or discarding entire groups of variables, making it suitable when group-level interpretability is desired. The Sparse Group Lasso combines the effects of both Elastic Net and Group Lasso, allowing for sparsity both at the group level and within groups. Finally, the Exclusive Lasso imposes a competitive structure within each group, encouraging the selection of at least one variable per group. This exclusivity within the group makes it particularly appealing in applications such as pathway-based genomic analysis, where it is often desirable to identify a single representative feature of each biologically significant group.

In addition to the methods discussed above, several other regularization techniques have been proposed to induce different sparsity patterns, particularly to accommodate grouping structures among variables, such as the Fused Lasso (Tibshirani et al., 2005) and the Smoothly Clipped Absolute Deviation (SCAD) penalty (Fan and Li, 2001). Notably, some methods were originally developed for variable selection in large-scale multi-omics studies, such as the Priority Lasso (Klau et al., 2018) and the Integrative Lasso with Penalty Factors (IPF-Lasso; Boulesteix et al., 2017). These approaches require specifying additional inputs, such as the prioritization of groups or penalty factors assigned to each group.

For example, the central idea behind IPF-Lasso is to apply Lasso regularization within each group while introducing group-specific penalty factors. These factors reflect the relative importance or weighting of each group and can either be specified a priori or selected through CV. The IPF-Lasso

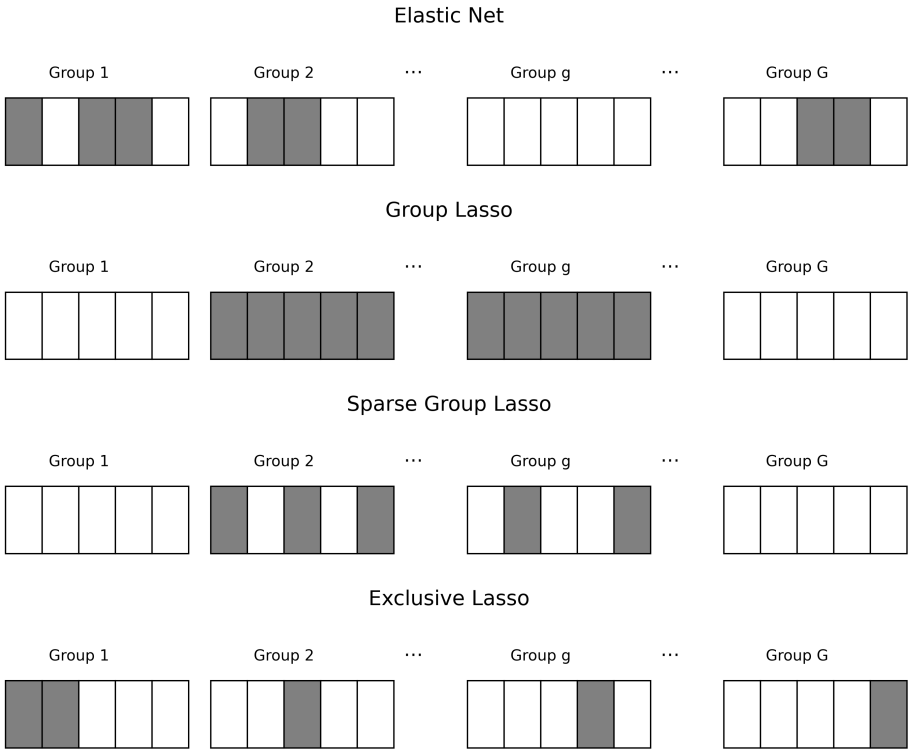

Figure 2.4: Sparsity patterns induced by different regularization methods. Each row shows the selected (grey) and excluded (white) coefficients across non-overlapping groups. Elastic Net enforces element-wise sparsity, Group Lasso selects entire groups, Sparse Group Lasso promotes both group-level and within-group sparsity, and Exclusive Lasso encourages within-group competition, selecting at least one feature per group.

penalty takes the form

$$P(\boldsymbol{\beta}) = \sum_{g \in \mathcal{G}} \lambda_g \|\boldsymbol{\beta}_g\|_1 \ ,$$

where λ_g denotes the penalty factor applied to group g, and $\boldsymbol{\beta}_g$ is the vector of coefficients for variables in that group. In practice, these group-specific penalties are typically chosen via CV by performing a grid search over a set of prespecified candidate vectors. However, this tuning process can be computationally intensive, particularly in high-dimensional settings.

2.3 Cox Proportional Hazards Model

In predictive modeling, apart from GLMs, time-to-event prediction problems require specialized statistical frameworks to handle censored data. Censoring occurs when the true event time for an observation is unknown, possibly due to reasons such as loss of follow-up or the end of the study period. The Cox Proportional Hazards (PH) model (Cox, 1972) is one of the most widely used techniques for time-to-event analysis, particularly in medical and biological research.

Let $i = 1, \ldots, n$ denote the observations (patients) in the cohort. For each patient, we observe the triplet $(t_i, \delta_i, \mathbf{x}_i)$, where t_i is the event or censoring time for patient i, δ_i is the censoring indicator (with $\delta_i = 1$ if the event is observed and $\delta_i = 0$ if the observation is censored), and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ is the vector of covariates.

The Cox PH model assumes that the hazard function for an individual with covariates $\mathbf{x}_i$ is given by:

$$h(t \mid \mathbf{x}_i) = h_0(t) \exp(\mathbf{x}_i^\top \boldsymbol{\beta}),$$

where $h_0(t)$ is the baseline hazard function, and $\boldsymbol{\beta}$ is the vector of regression coefficients. The hazard function $h(t \mid \mathbf{x}_i)$ represents the instantaneous risk of experiencing the event at time t, conditional on survival up to that time. Formally, it is defined as:

$$h(t \mid \mathbf{x}_i) = \lim_{\Delta t \to 0} \frac{\Pr(t \leq T < t + \Delta t \mid T \geq t, \mathbf{x}_i)}{\Delta t},$$

where T denotes the random variable representing the event time.

Unlike parametric survival models, the Cox PH model does not assume a specific form for the baseline hazard function $h_0(t)$, making it semi-parametric. This flexibility allows the model to estimate the effect of covariates without the need to specify the functional form of the baseline hazard. In practice, $h_0(t)$ can be estimated non-parametrically using techniques such as the Kaplan–Meier estimator for the survival function or the Nelson–Aalen estimator for the cumulative hazard.

To estimate $\boldsymbol{\beta}$, the Cox PH model utilizes the partial likelihood function, which emphasizes the relative risk of individuals experiencing an event rather than modeling the baseline hazard explicitly. The partial log-likelihood function for the Cox PH model is given by:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^\top \boldsymbol{\beta} - \log \left(\sum_{l \in R(t_i)} \exp(\mathbf{x}_l^\top \boldsymbol{\beta}) \right) \right], \tag{2.4}$$

where $R(t_i)$ represents the risk set at time t_i, which includes all individuals who are still at risk (i.e., uncensored and have not yet experienced the event) at the time of observation.

It is referred to as a partial likelihood because it excludes the baseline hazard function $h_0(t)$, relying solely on the ordering of event times rather than their exact values. This formulation simplifies the computation and enables the estimation of $\boldsymbol{\beta}$ without requiring the full likelihood function of the survival times.

In high-dimensional settings, variable selection is typically performed using Lasso regularization. In principle, any of the additional penalties discussed in the previous sections can be incorporated into the partial log-likelihood function of the Cox PH model, as defined in Equation (2.4). Numerous extensions of such penalized models have already been proposed in the literature (Tibshirani, 1997; Zhang and Lu, 2007; Kim et al., 2012).

However, to date, the Cox PH model with the Exclusive Lasso penalty has not been implemented. This extension is particularly important because, in most biological datasets, genes are organized into distinct biological pathways. Therefore, it is crucial to identify key genes in each pathway.

Moreover, since patient survival is often the clinical endpoint in these datasets, incorporating the Exclusive Lasso into the Cox PH framework becomes highly relevant.

In Article V of this thesis, we extend the Exclusive Lasso penalty, as defined in Equation (2.3), to the Cox PH model.

3. Summary of the Articles

This chapter outlines the contributions of all publications included in this cumulative dissertation. Since Article III represents the preliminary version of Article IV, the corresponding summaries are presented jointly.

3.1 Article I

Non-linear modelling of systolic and diastolic blood pressures via environmental factors

In this study, we employ GAMs to investigate the non-linear relationship between blood pressure and environmental variables, specifically temperature and relative humidity. We argue that traditional linear models cannot capture the complex interaction effects between these climatic factors, significantly influencing systolic (SBP) and diastolic blood pressure (DBP). To address this limitation, we incorporate bivariate tensor product splines within the GAM framework, enabling flexible and smooth modeling of non-linear associations among predictors. Our primary objective is to examine how environmental exposures affect blood pressure differently across warm and cold seasons.

The analysis is based on data from the SALIA cohort (Study on the Influence of Air Pollution on Lung, Inflammation, and Aging), comprising 635 older women assessed during two follow-up periods (2007–2008 and 2012–2013). Temperature and humidity data were collected up to 30 days before each clinical examination. In addition to assessing immediate exposure (lag 0), we evaluated various moving average lags (e.g., lag 0–10 and lag 0–30) to explore potential delayed effects. The models also adjusted for relevant socio-demographic and health-related covariates, including age, body mass index (BMI), smoking behavior, diabetes status, and urban versus rural residence. Subject-specific random effects were incorporated to account for repeated measures.

GAMs were fitted using identity link functions, with P-splines for univariate smooth terms and tensor product splines for the interaction between temperature and humidity. Data were stratified by season to compare patterns during warm months (April–September) and cold months (October–March). Model selection based on the Akaike Information Criterion (AIC; (Akaike, 1974)) indicated that lag 0 models best explained blood pressure during the warm season, whereas more extended lag structures

(e.g., lag 0–10 and lag 0–20) were more suitable for the cold season. Our findings indicate a strong, non-linear interaction between temperature and humidity, particularly during warmer months, with a p-value for the bivariate interaction below 0.01. In addition, we observed significant individual-level variability in blood pressure responses, with covariates such as age, BMI, and residential location playing notable roles. Notably, age exhibited a linear negative association with DBP.

In conclusion, our results highlight the importance of modeling interactions between climatic variables when examining blood pressure variation. Using GAMs with tensor product splines, we capture complex, seasonally dependent patterns that conventional models often overlook. While we did not perform comparisons with alternative interaction modeling approaches, our methodology nonetheless offers a robust and flexible framework for evaluating the environmental determinants of cardiovascular health.

3.2 Article II

Complex synergistic effects of air pollution and temperature on blood pressure: Evidence from the SALIA cohort study

In this article, we build upon the findings of Article I and address its limitations. In the previous study, we advocated the use of GAMs to model the non-linear interactions between environmental factors and blood pressure, presenting a proof of concept. However, we did not compare GAMs to alternative interaction modeling strategies or assess their relative advantages and drawbacks. To expand our analysis, we include additional covariates beyond temperature and relative humidity—specifically, three air pollutants: particulate matter ($PM_{2.5}$), nitrogen dioxide (NO_2), and ozone (O_3).

To investigate the potential synergistic effects of mean temperature (Tmean) and air pollution ($PM_{2.5}$, NO_2, and O_3) on systolic and diastolic blood pressure (SBP and DBP), we incorporate Tmean and pollutant concentrations into the model using three different approaches.

We begin with a simple linear interaction model to evaluate how Tmean modifies the effects of air pollution on SBP and DBP. In this model, Tmean is treated as a categorical variable, and the cutoffs are determined by the empirical distribution: "low" corresponds to temperatures below the 25th percentile, "medium" to the 25th—75th percentiles, and "high" to values above the 75th percentile.

To assess interaction effects, we include multiplicative terms between air pollutant concentrations and Tmean categories. We test the statistical significance of differences in pollutant effects across temperature strata (e.g., "low" or "high" Tmean vs. "medium" Tmean) by computing the 95% confidence interval for the difference in estimated effects:

$$(\hat{\beta}_1 - \hat{\beta}_2) \pm 1.96 \times \sqrt{\widehat{SE}_1^2 + \widehat{SE}_2^2}\,,$$

where $\hat{\beta}_1$ and $\hat{\beta}_2$ are the estimated pollutant effects at two Tmean categories, and $\widehat{SE}_1$ and $\widehat{SE}_2$ are their corresponding standard errors. This expression assumes independence of the estimates.

In the second model, we explore non-linear interaction effects by modeling pollutant concentrations as smooth functions within each Tmean category. This allows us to capture more flexible, temperature-specific associations. The third model utilizes bivariate tensor product penalized splines to jointly model Tmean and air pollutant concentrations as smooth interacting predictors. This approach allows for a fully non-linear representation of the joint exposure-response surface.

All models are specified as GAMs with an identity link function. We use P-splines for continuous covariates such as age, body mass index (BMI), and cigarette consumption (in pack-years). The specifications of the three models are as follows:

Model 1 (Linear interaction model):

$$log\ \mathbb{E}[\mathrm{Y}] = \mathrm{Tmean}_{category} \times \mathrm{PM}_{2.5}/\mathrm{NO}_2/\mathrm{O}_3 + \mathrm{Tmean} + [\mathrm{Covariates}]$$

Model 2 (Non-linear interaction model):

$$log\ \mathbb{E}[\mathrm{Y}] = s(\mathrm{PM}_{2.5}/\mathrm{NO}_2/\mathrm{O}_3\ ; \mathrm{Tmean}_{category}) + [\mathrm{Covariates}]$$

Model 3 (Tensor product model):

$$log\ \mathbb{E}[\mathrm{Y}] = te(\mathrm{Tmean}, \mathrm{PM}_{2.5}/\mathrm{NO}_2/\mathrm{O}_3) + [\mathrm{Covariates}]$$

Here, Y denotes blood pressure, $s(\cdot)$ denotes a univariate smooth function, and $te(\cdot, \cdot)$ is a bivariate tensor product spline capturing the interaction between Tmean and each air pollutant.

We evaluate the ability of these models to detect interactions between Tmean and air pollution on BP. Figure 3.1 displays the estimated change in SBP per unit increase in pollutant concentrations across Tmean categories under both linear and non-linear specifications. At higher Tmean levels, $\mathrm{PM}_{2.5}$ showed a positive association with SBP, with stronger effects at longer lag periods. At low and medium Tmean levels, linear models yielded inconclusive results. In contrast, the non-linear models revealed clear associations, indicating that most interactions between Tmean and air pollution were non-linear. As shown in Panel (c), the response surfaces obtained from Model 3, based on bivariate tensor product smooths, demonstrate that the association between air pollution and SBP is non-linear and influenced by the feature Tmean. Rather than contributing independently, Tmean and air pollution interact in a complex manner, influencing the overall pattern of their combined effect on SBP. Overall, the joint relationship between Tmean and pollutant levels appears as a smooth and non-linear surface.

Model 1, which only allows for linear interactions, failed to detect significant associations in most cases. While Model 2 captured non-linear effects, it may suffer from reduced statistical power due to the categorization of Tmean. Model 3, based on bivariate tensor product splines, effectively modeled the combined effects and allowed for greater flexibility in capturing complex interactions. Although it does not produce directly interpretable coefficients for interaction terms (Hastie and Tibshirani, 1990), it is statistically robust and accommodates differing measurement scales.

Furthermore, model performance was compared using the AIC, with Model 3 achieving the lowest AIC value. Therefore, we selected Model 3 as our core model.

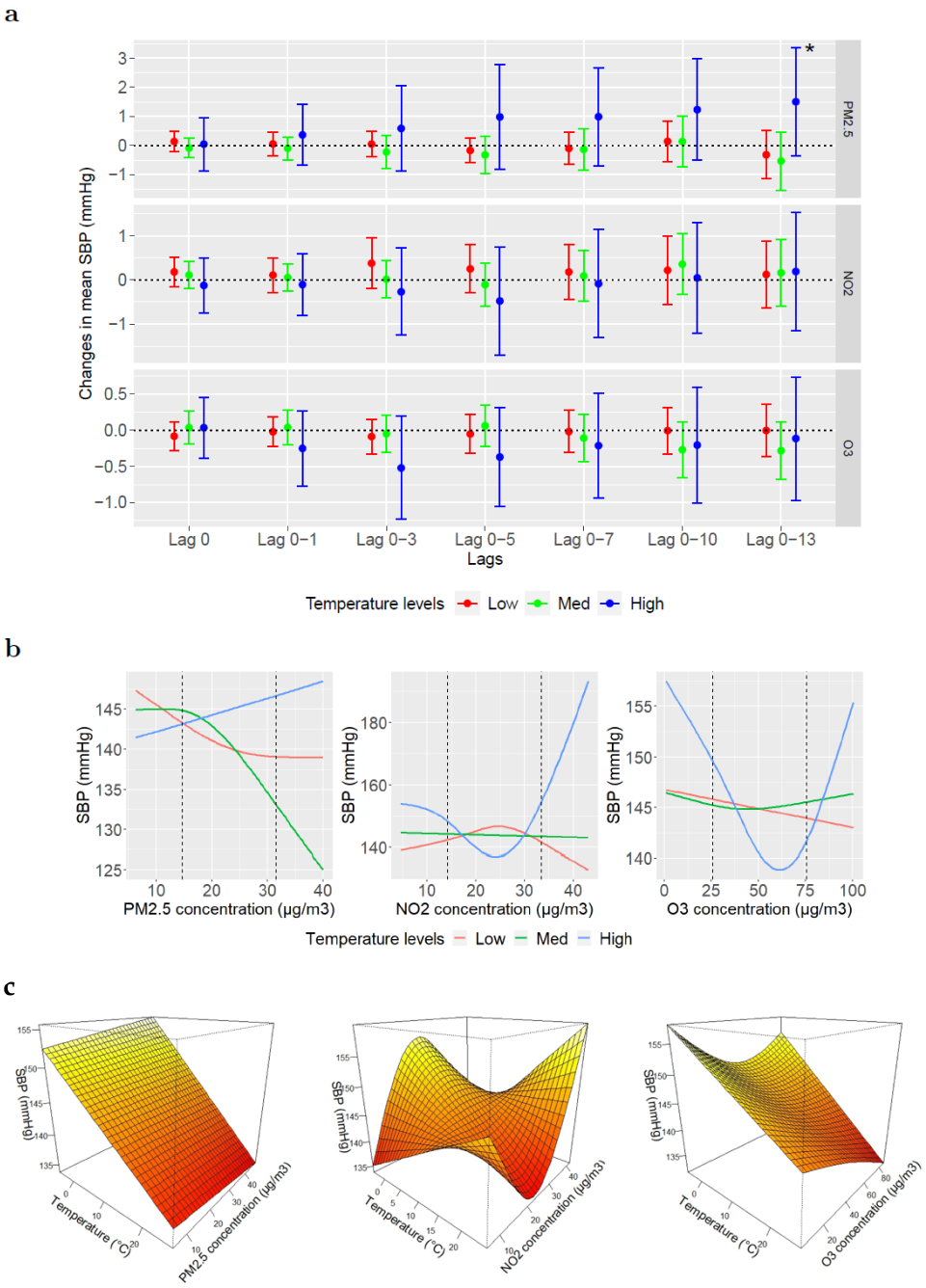

Figure 3.1: Panel (a): Estimated changes in systolic blood pressure (SBP, mmHg) per unit increase in $PM_{2.5}$, NO_2, and O_3 concentrations ($\mu g/m^3$) across temperature levels, based on Model 1. Panel (b): Adjusted non-linear associations between SBP and pollutant concentrations by temperature level, based on Model 2. Vertical dotted lines indicate the 25th, 50th, and 75th percentiles of pollutant exposure; an asterisk (*) marks a statistically significant difference between temperature levels. Panel (c): Bivariate response surfaces of temperature and air pollutants on SBP, based on Model 3. All models were adjusted for age, body mass index (BMI), socioeconomic status (SES), urban/rural location, cigarettes per day, second-hand smoke exposure, season, and relative humidity.

Focusing on short-term exposure, we found that low temperatures combined with elevated levels of $PM_{2.5}$ and NO_2 were associated with increased SBP and DBP. In contrast, at higher temperatures, both pollutants tended to lower SBP. O_3 consistently showed a negative association with blood pressure across all temperature levels, except at very low temperatures. These findings highlight the interactive effects of temperature and air pollution on blood pressure. Stratified analyses indicated that these effects were more pronounced in women residing in urban areas and those with a lower socioeconomic status. Overall, non-linear interactions were apparent, with blood pressure effects varying across temperature percentiles for each pollutant.

In conclusion, we demonstrate the utility of GAMs in modeling the combined effects of mean temperature and air pollutants on blood pressure. To make the article accessible to a broader audience in environmental epidemiology, we intentionally avoided technical modeling details and maintained a less statistically intensive focus. Nonetheless, the article emphasizes the importance of accounting for bivariate interactions and clearly illustrates how each air pollutant interacts with temperature to affect blood pressure. By advocating for this flexible and robust statistical framework, we provide a practical approach for advancing from single-pollutant to multi-pollutant analyses, addressing a key methodological gap in environmental epidemiology.

3.3 Articles III and IV

Optimizing Variable Selection in Multi-Omics Datasets: A Focus on Exclusive Lasso

This article explores the challenges of structured variable selection in high-dimensional biological data, particularly multi-omics datasets. Traditional Lasso fails to perform adequately in such contexts due to its inability to account for group structure and its weakness in handling highly correlated features within groups. We propose using the Exclusive Lasso, a regularization method based on the $L_{1,2}$-norm penalty to address this. This formulation induces sparsity within groups, promoting the selection of only one or a few representative variables from each group while allowing for flexibility between groups.

The proposed method, the *Newton Method $L_{1,2}$ (NM-$L_{1,2}$) Sparsity Algorithm*, employs a quadratic approximation to the non-differentiable L_1-norm, transforming the original problem into a differentiable one suitable for Newton-type optimization. The approach is evaluated through simulation studies under a range of correlation structures, showing substantial improvements in both variable selection accuracy and prediction error compared to standard Lasso.

The algorithm is then applied to a Down Syndrome classification task using a multi-omics dataset comprising methylation, glycomics, and clinical variables, with each omics layer treated as a separate group. The results demonstrate the proposed method's superior classification performance over both standard Lasso and coordinate-descent-based Exclusive Lasso. This work lays the foundation for further refinements to optimize the Exclusive Lasso, particularly in the context of structured biomedical data.

A Newton-based Variant of Exclusive Lasso for Improved Sparse Solutions

Building on the foundations of Article III, this article presents an enhanced version of the NM-$L_{1,2}$ Sparsity Algorithm by introducing an alternative smoothing approach that improves optimization stability in sparse, high-dimensional settings.

This article formally introduces the Exclusive Lasso problem and addresses its two primary limitations. Although it has been in the literature for some time, its applications remain relatively limited. The Exclusive Lasso is particularly suited for scenarios where variables within the same group compete against each other, for example, multi-omics or gene expression studies where features are naturally grouped by biological function or measurement platform.

In contrast to the Group Lasso (Yuan and Lin, 2006), which encourages the joint selection of all variables within a group, the Exclusive Lasso enforces intra-group sparsity through the L_1-norm and applies the L_2-norm across groups. This structure promotes competition among variables within the same group while maintaining representation from each group. The L_2-norm across groups prevents entire groups from being excluded by merely shrinking their coefficients, thereby increasing the likelihood of retaining at least one informative variable per group. However, two critical issues limit the standard Exclusive Lasso: (1) the L_1-norm induces a non-differentiable objective, making optimization difficult, and (2) it forces the selection of at least one variable from every group, even if some groups contain no useful information.

To address the first limitation, we employ smooth approximations to the L_1-norm. Building on the quadratic approximation introduced in Article III, we further propose a sigmoid-based approximation in this work. Both approaches transform the originally non-differentiable penalty term from Equation (2.3) into a smooth and continuous function, making it suitable for second-order optimization techniques. The approximated penalty functions are given as follows:

1. **Quadratic approximation:**

$$(2.3) \Rightarrow \frac{1}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} \sqrt{\beta_k^2 + c} \right)^2 .$$

2. **Sigmoid function approximation:**

$$(2.3) \Rightarrow \frac{1}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} \left[\left(\log \left(1 + e^{(-c\beta_k)} \right) + \log \left(1 + e^{(c\beta_k)} \right) \right) \right] \right)^2 .$$

Figure 3.2 shows both approximations. In the plot on the left (quadratic approximation), we observe a trade-off between smoothness and accuracy. A small value of c results in a closer match to $|x|$ but may make optimization unstable. Larger values of c make smoother functions easier to optimize but may distort the shape of $|x|$ near zero. For the sigmoid function, the behavior is reversed: a larger c gives a more accurate approximation. These approximations make the penalty term suitable for fast and efficient optimization, which is particularly important for high-dimensional biological datasets.

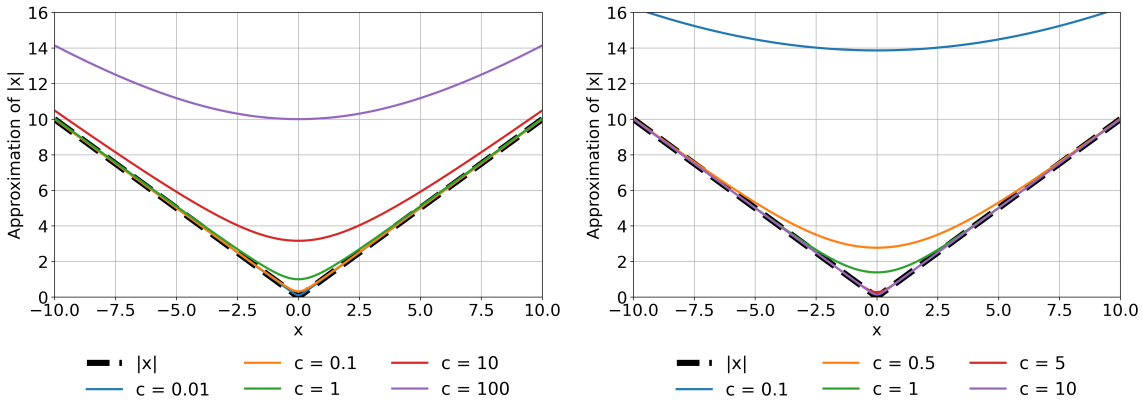

Figure 3.2: Approximations of the absolute value function $|x|$ using quadratic (*left*) and sigmoid-based (*right*) smoothing techniques for various values of the smoothing parameter c. The true $|x|$ function is shown as a black dashed line for reference.

To address the second limitation, we apply a small threshold to the final coefficient estimates. Any estimate below this threshold is automatically set to zero. This step not only mimics Lasso's behavior but also helps remove uninformative variables from all groups, eliminating the need to select at least one variable per group.

We perform a thorough simulation study to evaluate the performance of different smooth approximations under various settings. Consistent with findings in pure L_1-norm optimization (Schmidt et al., 2007), the sigmoid-based approximation generally converges in fewer iterations than the quadratic alternative. A similar pattern is observed for the Exclusive Lasso penalty, where the sigmoid approximation demonstrates superior computational efficiency. Furthermore, we explore the influence of the smoothing parameter c on approximation accuracy. Finally, we propose an enhanced NM-$L_{1,2}$ Sparsity Algorithm that addresses the Exclusive Lasso problem by utilizing adaptive step sizes and dynamically optimizing the parameter c in the sigmoid function approximation, resulting in improved computational efficiency.

To further test the model, we apply it to two real-world datasets. In the first case involving gene expression data, our model successfully selects informative genes across all modules, allowing for accurate classification of cancer-sensitive and resistant samples. In the second case, focused on financial data, our model achieves low prediction error while selecting fewer variables than the standard Lasso.

Articles III and IV together present a comprehensive methodological framework for addressing the Exclusive Lasso problem in practice. The proposed variant, optimized via the NM-$L_{1,2}$ algorithm, offers an efficient and scalable solution that is particularly well-suited for high-dimensional, group-structured datasets commonly encountered in domains such as multi-omics and finance.

3.4 Article V

Time-to-event prediction for grouped variables using Exclusive Lasso

Motivated by the increasing use of multi-omics data in survival modeling and the challenges posed by their complex structure, we investigated regularization techniques that account for group-specific information. These datasets, which integrate various types of high-dimensional molecular data, have become increasingly important in biomedical research, particularly for improving the prediction of patient survival (Hasin et al., 2017). Traditional approaches often rely on a single omics type or treat all features equally, which can result in omitting important low-dimensional groups such as clinical variables (Boulesteix and Sauerbrei, 2011). Recent studies have shown that incorporating group structure may improve model performance and help retain key predictors from smaller but informative groups like clinical data (Herrmann et al., 2021). Based on these insights, we developed a Cox PH model with Exclusive Lasso regularization to support structured feature selection across multiple data types.

The manuscript begins by reviewing regularization methods incorporating group structures among predictors, such as Elastic Net, Sparse Group Lasso, and IPF-Lasso. While these methods have been adapted for time-to-event modeling and prediction, they do not ensure that each group is represented in the final model. In contrast, the Exclusive Lasso explicitly encourages the selection of at least one variable from each group. This is particularly important in multi-omics settings, where conventional methods tend to select highly correlated features and may overlook informative variables from smaller or less correlated groups. By applying Exclusive Lasso, it is possible to retain important signals from low-dimensional groups, such as clinical variables, which have been shown to improve survival prediction in multi-omics data significantly (Herrmann et al., 2021).

We add the penalty of Exclusive Lasso from Equation (2.3) to the partial log-likelihood function for the Cox PH model from Equation (2.4). The estimation is then carried out by maximizing the penalized partial log-likelihood function, given by

$$\begin{aligned}\ell_{\text{pen}}(\boldsymbol{\beta}) &= \ell(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta}) \\ &= \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \log \left(\sum_{l \in R(t_i)} \exp(\mathbf{x}_l^T \boldsymbol{\beta}) \right) \right] - \frac{\lambda}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} |\beta_k| \right)^2 ,\end{aligned}$$

where $\lambda \geq 0$ is the penalty parameter.

We introduce a coordinate descent algorithm analogous to that used in standard Exclusive Lasso. The gradient of the Cox partial likelihood is well-known, and we incorporate soft-thresholding to manage the L_1-norm component of the penalty. Soft-thresholding is widely used in the literature as it enables efficient optimization in sparse models like the Lasso by shrinking coefficients towards zero and setting small ones exactly to zero, thereby achieving automatic variable selection. We demonstrate that the coordinate descent procedure ensures convergence for the proposed method.

We perform extensive simulation studies to benchmark Exclusive Lasso against Elastic Net, Group Lasso, and IPF-Lasso under various scenarios with differing group sizes and numbers of true signal

variables. The results show that Exclusive Lasso consistently attains the highest selection accuracy and lowest false discovery rate, primarily when signals are evenly distributed across groups. Although its performance slightly diminishes under randomly allocated signals, it remains competitive and often outperforms IPF-Lasso, the next best-performing approach.

The method's real-world utility is demonstrated by applying it to a bladder cancer dataset from the GEO database. This dataset includes both gene expression and clinical variables. Using a stability selection-inspired procedure (Meinshausen and Bühlmann, 2010), we identify the top ten most frequently selected features across 100 iterations. Notably, Exclusive Lasso consistently selects at least one variable from the low-dimensional clinical group in every iteration, while other methods display instability, selecting different variables across runs. The improved predictive performance is attributed to Exclusive Lasso's ability to identify key variables across all groups, regardless of their dimensionality.

In summary, this paper presents Exclusive Lasso as a compelling and practical tool for survival modeling and prediction in grouped high-dimensional data. Its capacity to ensure representation from all relevant variable groups addresses a significant limitation of traditional penalized regression models.

3.5 Software

For Articles I and II, the analyses were conducted using the `mgcv` package in R (Wood, 2017, 2011). In particular, we employed the `te()` function with `bs="ps"` to specify P-splines, and model estimation was carried out primarily using the "REML" method.

For Articles III and IV, all computations were newly implemented from scratch in Python (version 3.12.3). The NM-$L_{1,2}$ sparsity algorithm was coded and integrated into the analysis pipeline. The complete source code is publicly available on GitHub at: `https://github.com/draviis/NML12.git`.

For Article V, the original R implementation of the Exclusive Lasso package (Weylandt et al., 2018) was extended to support Cox PH models. The source code is available on GitHub at: `https://github.com/draviis/ExclusiveLassoCox`

4. Conclusion, Discussion and Outlook

This cumulative thesis brings together five manuscripts that address three key challenges in modeling environmental and biomedical datasets. While predictors in these datasets are often modeled as having straightforward linear effects, their true influence is typically more complex due to interactions with other variables. Moreover, the high dimensionality of such data makes it challenging to identify the relevant predictors, especially when variables exhibit intricate correlation structures.

The first study applies GAMs to capture complex interactions between environmental factors. Although GAMs are widely used in environmental research due to their flexibility in modeling non-linear relationships between predictors and outcomes (Jbilou and El Adlouni, 2012), they are most commonly applied in single-pollutant studies. Given the growing need for a multi-pollutant perspective, accounting for correlations among atmospheric variables is essential. This study addresses this gap by modeling the joint effects of environmental predictors using bivariate tensor product interactions, thereby tackling the first key challenge of capturing complex relationships between environmental exposures.

Two articles contribute to this part of the thesis. In Article I, we examine the interaction between temperature and relative humidity and its effect on the blood pressure of elderly German women using a bivariate tensor product smooth. We stratified the data by season to investigate whether the interactions vary between warm and cold months, as previous studies suggest that seasonal variation influences blood pressure. For each season, we fit a separate GAM, which provides a clearer picture of the seasonal patterns in the non-linear interactions. This stratification is justified by our findings, which indicate that the interaction between temperature and humidity plays a critical role in blood pressure modeling during warmer months (p-values < 0.01), whereas the interaction is not statistically significant in colder months. Furthermore, we observe that the effect of the interaction appears to be immediate in warm months, while in colder months the influence emerges with a time lag. These findings are consistent with earlier epidemiological studies (Brook, 2017).

Article II extends this approach by investigating the combined effects of temperature and three air pollutants on systolic and diastolic blood pressure. Previous studies examine the short-term effects of temperature and air pollution on blood pressure, often assuming linear relationships and reporting mixed results: some show positive associations (Ishii et al., 2020; Wen et al., 2023), while others find negative associations (Chen et al., 2013; Giorgini et al., 2015). One possible reason for these inconsistencies is the exclusion of synergistic interactions among environmental exposures

and the assumption of linear effects. Our study addresses both issues by modeling potential non-linearities and interactions. We find that the association between air pollution and blood pressure varies with temperature, indicating strong interactive effects. For example, we observe an increase in blood pressure associated with $PM_{2.5}$ exposure at low-temperature levels. In contrast, $PM_{2.5}$ exposure is linked to a decrease in blood pressure at medium and high-temperature levels. Stratified analyses also identify particularly vulnerable subgroups, such as older women from socioeconomically disadvantaged backgrounds. The study benefits significantly from the well-characterized SALIA cohort (Schikowski et al., 2005), which includes detailed information on confounders such as smoking and alcohol consumption, thereby allowing for confounding adjustment. Our findings indicate that a combination of low ambient temperature and high air pollution levels is associated with a substantial increase in blood pressure among elderly German women.

The second study focuses on the key challenge of variable selection. As observed in the previous study, variables often interact with one another and are frequently correlated. Therefore, it is critical to account for these relationships when selecting variables for predictive modeling. A further complication arises from the presence of many potential influence factors, which can often be naturally grouped based on their origin or nature. To address this, we propose the use of Exclusive Lasso, a regularization method designed to select variables across predefined groups while ensuring that no group is entirely excluded. However, optimizing Exclusive Lasso presents challenges due to the non-differentiability of the L_1-norm. To overcome this, we follow the approaches of Schmidt et al. (2007) and Oelker and Tutz (2017), which provide a differentiable approximation of the L_1-norm. Building on this idea, Article III introduces a quadratic approximation to the Exclusive Lasso penalty, resulting in a smooth formulation that enables efficient gradient-based optimization.

In Article IV, we extend this methodology by adding another approximation and proposing a new algorithm that improves on the standard Exclusive Lasso, particularly in scenarios where representation from every group is not required. Our variant includes a thresholding step, which allows the model to discard non-informative groups more effectively. This approach leads to improved overall performance compared to the original method. The corresponding algorithm is implemented in Python, and the source code is publicly available on GitHub.

The final study focuses on improving time-to-event prediction in biomedical datasets such as multi-omics, where low-dimensional variables like clinical features are often overlooked. We extend the Exclusive Lasso framework to the Cox PH model to address this. Rather than using our custom optimization technique from Article IV, we adopt a simpler approach based on the coordinate descent algorithm. We extend the basic algorithm implemented in the original `ExclusiveLasso` R package (Campbell and Allen, 2017; Weylandt et al., 2018) to the Cox PH setting. This strategy aligns well with our aim to ensure representation from each variable group, including smaller ones like clinical variables. The coordinate descent algorithm updates one coefficient at a time and tends to select a representative from each group early in the optimization process. Additional variables are added as needed, which helps preserve group-specific information. Coordinate descent also performs well for Cox PH models and is widely used in established packages such as `glmnet` (Simon et al., 2011). Using warm starts, which initialize the algorithm with the solution from previous runs, significantly improves convergence speed and makes the method highly suitable for high-dimensional data. We evaluate the proposed method on a real-world cancer dataset and compare its performance with

state-of-the-art regularization techniques incorporating group structures. The results show that Exclusive Lasso consistently selects clinical variables across iterations, a property not observed in competing methods.

The following outlines the general limitations of the thesis and proposes directions for future research. In this work, we addressed the challenge of variable selection in GLMs and Cox PH models, particularly in the presence of high correlations among predictors. However, as demonstrated in our first study, the influence of particular variables may be non-linear and better captured using flexible modeling techniques such as GAMs. This raises the need for effective variable selection methods explicitly tailored to GAMs. A key complication in GAMs is *concurvity*, which refers to the non-linear dependence between predictors and can lead to unstable parameter estimates, analogous to multicollinearity in linear models (Ramsay et al., 2003). Kovács (2024) reviews feature selection techniques developed for GAMs. For instance, GAMBoost, a component-wise boosting method for additive models, allows automatic effect selection, allowing each covariate to enter the model either linearly or non-linearly or be excluded entirely (Hothorn et al., 2010; Groll and Tutz, 2012). Incorporating Lasso-like penalties into GAMs is also possible by applying regularization to the coefficients of the basis functions. Specifically, an L_1-norm penalty on the basis function coefficients can shrink all coefficients of a feature to zero, effectively removing the feature from the model (Marra and Wood, 2011). To extend this approach to grouped variables, algorithms based on the Hilbert–Schmidt Independence Criterion (HSIC) have been proposed. These methods assess pairwise independence between features and have already been implemented in the Group Lasso setting (Yamada et al., 2014). This makes the extension to the Exclusive Lasso framework a natural next step.

Extending the proposed techniques to the GAM framework, particularly in high-dimensional settings, may present computational challenges. Our current approach uses a standard Newton–Raphson algorithm to optimize the objective function. However, for high-dimensional problems where the computation of the Hessian matrix becomes impractical, we recommend using Quasi-Newton methods, which approximate the Hessian using gradient information from successive iterations. Another promising alternative is the Penalized Iteratively Reweighted Least Squares (PIRLS) method, as proposed by Oelker and Tutz (2017), which may offer substantial improvements in computational efficiency. In addition, different smoothing strategies for the composite Exclusive Lasso penalty could be explored. For instance, reformulating the problem as a quadratic objective with linear constraints or applying Least Absolute Deviation (LAD) regression (Steiger et al., 1983), which can be expressed as a linear programming problem, could enable the use of efficient optimization techniques within the Exclusive Lasso framework. This idea may be extended to incorporate Elastic Net penalization, which could provide more structured sparsity and be particularly beneficial in the presence of highly correlated predictors.

Bibliography

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6):716–723.

Binder, H. and Schumacher, M. (2008). Allowing for mandatory covariates in boosting estimation of sparse high-dimensional survival models. *BMC Bioinformatics*, 9:1–10.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative l1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and Mathematical Methods in Medicine*, 2017(1):7691937.

Boulesteix, A.-L. and Sauerbrei, W. (2011). Added predictive value of high-throughput molecular data to clinical data and its validation. *Briefings in Bioinformatics*, 12(3):215–229.

Bøvelstad, H. M., Nygård, S., and Borgan, Ø. (2009). Survival prediction from clinico-genomic models-A comparative study. *BMC Bioinformatics*, 10:1–9.

Breheny, P. and Huang, J. (2015). Group descent algorithms for nonconvex penalized linear and logistic regression models with grouped predictors. *Statistics and Computing*, 25(2):173–187.

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (2017). *Classification and regression trees*. Routledge.

Brook, R. D. (2017). The environment and blood pressure. *Cardiology clinics*, 35(2):213–221.

Campbell, F. and Allen, G. I. (2017). Within group variable selection through the exclusive lasso. *Electronic Journal of Statistics*, 11:4220–4257.

Chen, G., Song, G., Jiang, L., Zhang, Y., Zhao, N., Chen, B., and Kan, H. (2008). Short-term effects of ambient gaseous pollutants and particulate matter on daily mortality in Shanghai, China. *Journal of Occupational Health*, 50(1):41–47.

Chen, Q., Wang, J., Tian, J., Tang, X., Yu, C., Marshall, R. J., Chen, D., Cao, W., Zhan, S., Lv, J., et al. (2013). Association between ambient temperature and blood pressure and blood pressure regulators: 1831 hypertensive patients followed up for three years. *PloS one*, 8(12):e84522.

Cheng, Y. and Kan, H. (2012). Effect of the interaction between outdoor air pollution and extreme temperature on daily mortality in Shanghai, China. *Journal of Epidemiology*, 22(1):28–36.

Cox (1972). Regression models and life tables. *JR Stat Soc*, 34:248–275.

De Boor, C. (1972). On calculating with B-splines. *Journal of Approximation theory*, 6(1):50–62.

Dominici, F., Peng, R. D., Barr, C. D., and Bell, M. L. (2010). Protecting human health from air pollution: shifting from a single-pollutant to a multipollutant approach. *Epidemiology*, 21(2):187–194.

Eilers, P. H. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.

Eilers, P. H. and Marx, B. D. (2021). *Practical smoothing: The joys of P-splines*. Cambridge University Press.

Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association*, 96(456):1348–1360.

Giorgini, P., Rubenfire, M., Das, R., Gracik, T., Wang, L., Morishita, M., Bard, R. L., Jackson, E. A., Fitzner, C. A., Ferri, C., et al. (2015). Particulate matter air pollution and ambient temperature: opposing effects on blood pressure in high-risk cardiac patients. *Journal of Hypertension*, 33(10):2032–2038.

Groll, A. and Tutz, G. (2012). Regularization for generalized additive mixed models by likelihood-based boosting. *Methods of Information in Medicine*, 51(02):168–177.

Hasin, Y., Seldin, M., and Lusis, A. (2017). Multi-omics approaches to disease. *Genome biology*, 18:1–15.

Hastie, T. and Tibshirani, R. (1986). Generalized additive models. *Statistical Science*, pages 297–310.

Hastie, T. and Tibshirani, R. (1990). Exploring the nature of covariate effects in the proportional hazards model. *Biometrics*, pages 1005–1016.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):167.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based boosting 2.0. *Journal of Machine Learning Research*, pages 1851–1855.

Huang, Y. and Liu, J. (2018). Exclusive sparsity norm minimization with random groups via cone projection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6145–6153.

Ishii, M., Seki, T., Sakamoto, K., Kaikita, K., Miyamoto, Y., Tsujita, K., Masuda, I., and Kawakami, K. (2020). Association of short term exposure to asian dust with increased blood pressure. *Scientific Reports*, 10(1):17630.

Jbilou, J. and El Adlouni, S. (2012). Generalized additive models in environmental health: A literature review. *Novel approaches and their applications in risk assessment*, 120:2014–2016.

Kan, H., Chen, B., Zhao, N., London, S. J., Song, G., Chen, G., Zhang, Y., Jiang, L., et al. (2010). Part 1. a time-series study of ambient air pollution and daily mortality in Shanghai, China. *Research report (Health Effects Institute)*, (154):17–78.

Kim, J., Sohn, I., Jung, S.-H., Kim, S., and Park, C. (2012). Analysis of survival data with group lasso. *Communications in Statistics-Simulation and Computation*, 41(9):1593–1605.

Klau, S., Jurinovic, V., Hornung, R., Herold, T., and Boulesteix, A.-L. (2018). Priority-lasso: a simple hierarchical approach to the prediction of clinical outcome using multi-omics data. *BMC Bioinformatics*, 19:1–14.

Kong, D., Fujimaki, R., Liu, J., Nie, F., and Ding, C. (2014). Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. *Advances in Neural Information Processing Systems*, 27.

Kovács, L. (2024). Feature selection algorithms in generalized additive models under concurvity. *Computational Statistics*, 39(2):461–493.

Li, W., Pei, L., Li, A., Luo, K., Cao, Y., Li, R., and Xu, Q. (2019). Spatial variation in the effects of air pollution on cardiovascular mortality in Beijing, China. *Environmental Science and Pollution Research*, 26:2501–2511.

Li, Y., Xiao, C., Li, J., Tang, J., Geng, X., Cui, L., and Zhai, J. (2018). Association between air pollution and upper respiratory tract infection in hospital outpatients aged 0–14 years in Hefei, China: A time series study. *Public health*, 156:92–100.

Lin, M., Yuan, Y., Sun, D., and Toh, K.-C. (2020). Adaptive sieving with ppdna: Generating solution paths of exclusive lasso models. *arXiv preprint arXiv:2009.08719*.

Marra, G. and Wood, S. N. (2011). Practical variable selection for generalized additive models. *Computational Statistics & Data Analysis*, 55(7):2372–2387.

Mauderly, J. L., Burnett, R. T., Castillejos, M., Özkaynak, H., Samet, J. M., Stieb, D. M., Vedal, S., and Wyzga, R. E. (2010). Is the air pollution health research community prepared to support a multipollutant air quality management framework? *Inhalation toxicology*, 22(sup1):1–19.

Mauderly, J. L. and Samet, J. M. (2009). Is there evidence for synergy among air pollutants in causing health effects? *Environmental health perspectives*, 117(1):1–6.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Nkansah, H., Benyah, F., and Amankwah, H. (2021). Smoothing approximations for least squares minimization with l1-norm regularization functional. *International Journal of Analysis and Applications*, 19(2):264–279.

Oelker, M.-R. and Tutz, G. (2017). A uniform framework for the combination of penalties in generalized structured models. *Advances in Data Analysis and Classification*, 11:97–120.

Pearce, J. L., Beringer, J., Nicholls, N., Hyndman, R. J., and Tapper, N. J. (2011). Quantifying the influence of local meteorology on air quality using generalized additive models. *Atmospheric Environment*, 45(6):1328–1336.

Ramsay, T. O., Burnett, R. T., and Krewski, D. (2003). The effect of concurvity in generalized additive models linking mortality to ambient particulate matter. *Epidemiology*, 14(1):18–23.

Ravindra, K., Rattan, P., Mor, S., and Aggarwal, A. N. (2019). Generalized additive models: Building evidence of air pollution, climate change and human health. *Environment International*, 132:104987.

Royston, P., Altman, D. G., and Sauerbrei, W. (2006). Dichotomizing continuous predictors in multiple regression: a bad idea. *Statistics in Medicine*, 25(1):127–141.

Schikowski, T., Sugiri, D., Ranft, U., Gehring, U., Heinrich, J., Wichmann, H.-E., and Krämer, U. (2005). Long-term air pollution exposure and living close to busy roads are associated with copd in women. *Respiratory Research*, 6:1–10.

Schmidt, M., Fung, G., and Rosales, R. (2007). Fast optimization methods for l1 regularization: A comparative study and two new approaches. In *Machine Learning: ECML 2007: 18th European Conference on Machine Learning, Warsaw, Poland, September 17-21, 2007. Proceedings 18*. Springer. 286–297.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2019). *SGL: Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization*. R package version 1.3.

Simon, N., Friedman, J., Tibshirani, R., and Hastie, T. (2011). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Steiger, P. B. W. et al. (1983). Least absolute deviations. theory, applications, and algorithms.

Sun, Y., Chain, B., Kaski, S., and Shawe-Taylor, J. (2020). Correlated feature selection with extended exclusive group lasso. *arXiv preprint arXiv:2002.12460*.

Sun, Z., Tao, Y., Li, S., Ferguson, K. K., Meeker, J. D., Park, S. K., Batterman, S. A., and Mukherjee, B. (2013). Statistical strategies for constructing health risk models with multiple pollutants and their interactions: possible choices and comparisons. *Environmental Health*, 12:1–19.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 58(1):267–288.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(1):91–108.

Wen, T., Liao, D., Wellenius, G. A., Whitsel, E. A., Margolis, H. G., Tinker, L. F., Stewart, J. D., Kong, L., and Yanosky, J. D. (2023). Short-term air pollution levels and blood pressure in older women. *Epidemiology*, 34(2):271–281.

Weylandt, M., Campbell, F., and Allen, G. (2018). *ExclusiveLasso: Generalized Linear Models with the Exclusive Lasso Penalty*. R package version 0.0.

Wood, S. N. (2006). Low-rank scale-invariant tensor product smooths for generalized additive mixed models. *Biometrics*, 62(4):1025–1036.

Wood, S. N. (2011). Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *Journal of the Royal Statistical Society (B)*, 73(1):3–36.

Wood, S. N. (2017). *Generalized additive models: an introduction with R*. Chapman and Hall/CRC.

Yamada, M., Jitkrittum, W., Sigal, L., Xing, E. P., and Sugiyama, M. (2014). High-dimensional feature selection by feature-wise kernelized lasso. *Neural computation*, 26(1):185–207.

Yamada, M., Koh, T., Iwata, T., Shawe-Taylor, J., and Kaski, S. (2017). Localized lasso for high-dimensional regression. In *Artificial Intelligence and Statistics*, pages 325–333. PMLR.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhang, H. H. and Lu, W. (2007). Adaptive lasso for Cox's proportional hazards model. *Biometrika*, 94(3):691–703.

Zhang, T., Ghanem, B., Liu, S., Xu, C., and Ahuja, N. (2015). Robust visual tracking via exclusive context modeling. *IEEE transactions on Cybernetics*, 46(1):51–63.

Zhou, Y., Jin, R., and Hoi, S. C.-H. (2010). Exclusive lasso for multi-task feature selection. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 988–995. JMLR Workshop and Conference Proceedings.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

Part II

Publications

Article I

Non-linear modelling of systolic and diastolic blood pressures via environmental factors

Dayasri Ravi[1], Andreas Groll[1], Tamara Schikowski[2]

[1] Department of Statistics, TU Dortmund University, Germany
[2] IUF-Leibniz Research Institute for Environmental Medicine, Germany

E-mail for correspondence: ravi@statistik.tu-dortmund.de

Abstract: Systolic and diastolic blood pressures have always been closely associated with environmental factors such as temperature and relative humidity. However, the interaction effect between these environmental factors in modelling blood pressure is often not considered. We aim to use generalized additive models to model blood pressures as the environmental data often display a non-linear pattern. The explanatory variables may often have different measuring units. The tensor product spline approach is practical to model the interaction effect among the environmental explanatory variables instead of the isotropic smoothing.

Keywords: Generalized additive models; Tensor product splines; P-splines.

1 Introduction

Several studies have shown a significant relationship between blood pressure, temperature and relative humidity (e.g., Barnett et al., 2007). Although blood pressures are predicted reasonably well using environmental variables such as temperature, it is interesting to consider the interaction among these variables. We see that Generalized additive models (GAMs) are a better choice over Generalized linear models (GLMs) for modelling the effects of climatic and environmental variables (see Ravindra et al., 2019). Additive models are the sum of smooth, typically non-linear functions of the explanatory variables. These models allow for more flexible relationship between the variables compared to linear modelling. As the nature of the relationship between the response and explanatory variable(s) dictates the model, it can be analyzed non-parametrically. GAMs can be seen as an extension of GLMs, allowing the linear predictor to be expressed as smooth additive functions. GAMs can thus handle both non-linear and non-

This paper was published as a part of the proceedings of the 36th International Workshop on Statistical Modelling (IWSM), Trieste, Italy, 18–22 July 2022.

monotonous relationships between the response and explanatory variables. The variety of choices in smoothing functions can substantially improve the performance of the GAM model.
We analyzed the systolic and diastolic blood pressures of 635 women examined during two follow-ups. The aim was to model the blood pressures via socio-demographic covariates such as age, household, smoking pattern, schooling years, etc., and focus on the interaction effect between the environmental covariates.

2 Materials, Methods and Data

In the following section, we introduce the idea of GAMs and bivariate tensor product splines to model interaction effects. We briefly explain the data and the statistical analyses used. Finally, we state the critical results found in our study.

2.1 Methodology

Consider response variable y_i associated to covariates $\{x_{1i}, x_{2i}, \ldots, x_{pi}\}$. The GAM for modelling the data $\{y_i, x_{1i}, x_{2i}, \ldots, x_{pi}; i = 1, 2, .., n\}$ can be extended (see Wood, S. N., 2006) from GLMs with $g(\cdot)$ as a known monotonic differentiable link function to have the following structure

$$g(E[y_i]) = \beta_0 + f_1(x_{1i}) + f_2(x_{2i}) + \ldots + f_p(x_{pi}) \ . \tag{1}$$

Here, the functions $f_1(\cdot), f_2(\cdot), \ldots, f_p(\cdot)$ are unknown smooth functions to be estimated. To simplify the estimation of these functions, we represent them in such a way that Equ. (1) becomes a linear model. For simplicity, in the following we consider a single function $f(x)$ of one covariate x. We can represent the function $f(\cdot)$ through a linear combination of a set of basis functions $\{b_j; j = 1, 2, \ldots, d\}$ as

$$f(x) = \sum_{j=1}^{d} \alpha_j b_j(x) \ , \tag{2}$$

where the α_j's are unknown spline coefficient parameters and d is the corresponding number of basis functions.
In the following, we use the B-spline basis function approach for the basis functions $b_j(x)$ (see Eilers, P. H., and Marx, B. D., 2021). The smoothness and the number of B-splines depend on the number of knots. So, we consider a roughness penalty to overcome this strong dependence on the number of knots. Penalization can be applied to k-th order differences between the coefficients α_j from Equ. (2). Thus, instead of regular least squares, i.e. $\sum_{i=1}^{n}(y_i - f(x_i))^2$, we minimize the penalized least squares (PLS) criterion.

This is called the penalized spline (P-splines; see Eilers, P. and Marx, B., 2021) approach.
The k-th order P-spline based on $d = l + m - 1$ B-splines can be estimated by the following penalized residual sum of squares

$$PLS(\lambda) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{d}\alpha_j b_j(x_i))^2 + \lambda \sum_{j=k+1}^{d}(\Delta^k \alpha_j)^2, \tag{3}$$

where Δ^k is the k-th order difference among the coefficients α_j. $\lambda > 0$ is known as the smoothing parameter that controls the trade-off between the model fitting and smoothness.
In order to model the interaction between two covariates, x_1 and x_2, we can use the tensor product bases. We construct the univariate bases for x_1 and x_2 via $a_j(x_1), j = 1, 2, \ldots, d_1$, and $b_k(x_2), k = 1, 2, \ldots, d_2$, respectively. The bivariate smooth function $f_{12}(x_1, x_2)$ of x_1, x_2 has the following form

$$f_{12}(x_1, x_2) = \sum_{j=1}^{d_1}\sum_{k=1}^{d_2}\gamma_{jk} a_j(x_1) b_k(x_2) \ . \tag{4}$$

Here, we apply double penalization to both rows and columns of B-splines with respective smoothing parameters λ_1 and λ_2.
In case that random effects are included in the model, we can extend the predictor from Equ. (1) by $\mathbf{Zb}$, where $\mathbf{b} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}\sigma_b^2)$ represents the vector of random effects and $\mathbf{I}$ an identity matrix of suitable dimension with respect to the random effects components.

2.2 Study population

Data were collected from the IUF-Leibniz Research Institute for Environmental Medicine, as a part of the Study on Influence of Air Pollution on Lung, Inflammation and Aging (SALIA) cohort. The data comprises 635 women examined for systolic and diastolic blood pressures on two follow-ups recorded between 2007 to 2008 and 2012 to 2013. The climate data, temperature (in °C) and relative humidity (in %), were collected up to 30 days before the examination. Additional socio-demographic covariates such as age, Body Mass Index (BMI), years in school, smoking behaviour, diabetes, living conditions, location, etc., were updated at each follow-up.

2.3 Statistical analyses

We fit the blood pressure to a GAM with an identity link, and we select P-splines as the smoothers for the covariates. We model the temperature and relative humidity interaction via a bivariate tensor product P-spline. Additionally, we consider random intercepts for each study participant to

account for subject-specific variability. Taking the seasonal variations into consideration, we subset the data into warmer and colder months based on the date of examination and temperature. We examine the delayed effects of environmental variables by taking moving average lags up to 30 days before the examination. The lag structure includes lag 0, lag 0-1, lag 0-3, lag 0-5, lag 0-10, lag 0-20 and lag 0-30. Here, lag 0 represents the daily mean temperature and daily relative humidity taken on the day of examination (`temp_0`,`rh_0`), i.e. no lag at all, and lag 0-30 represents the moving average of daily mean temperature and daily relative humidity taken from the day of examination to 30 days before the examination (`temp_mean30`,`rh_mean30`). We choose the best moving average lag structure based on the Akaike information criterion (AIC).

2.4 Results

We find a huge subject-specific random effect for both systolic and diastolic blood pressures. The estimated standard deviations for the random effects are 15.289 (12.807-18.252, 95 % CI) and 7.579 (6.298-9.121, 95 % CI) for systolic and diastolic blood pressures, respectively. Based on the AIC score, we select the model with lag 0 for systolic and diastolic blood pressures for the warmer months (April - September). During colder months (October - March), we select the model with lag 0-10 for systolic blood pressure and lag 0-20 for diastolic blood pressure. Fig. 1 displays a robust and highly non-linear interaction effect between daily mean temperatures and daily relative humidity. From Fig. 2, we observe that the effect of age is linear for both systolic and diastolic blood pressures. We also notice a negative effect of age on diastolic blood pressure, which seems to be consistent with epidemiological results (e.g., Pinto E., 2007).
Our findings suggest that the bivariate interaction between temperature and relative humidity plays a critical role in modelling the blood pressures during the warmer months (p-values < 0.01) more than in colder months, where the effect was not significant. During colder months, the covariate location (rural or urban) shows significance (p-value $= 0.00674$ for diastolic and p-value $= 0.0099$ for systolic). Age also plays a significant role (p-value < 0.01) for all models except for systolic blood pressure in colder months. Some covariates such as the number of smoking packets per day and BMI show a non-linear effect on blood pressure. However, other covariates, such as years in school, diabetes, and heating conditions do not seem relevant for modelling systolic and diastolic blood pressures.

3 Conclusion

We have proposed a generalized additive model to understand the effect of environmental factors on systolic and diastolic blood pressures. We also

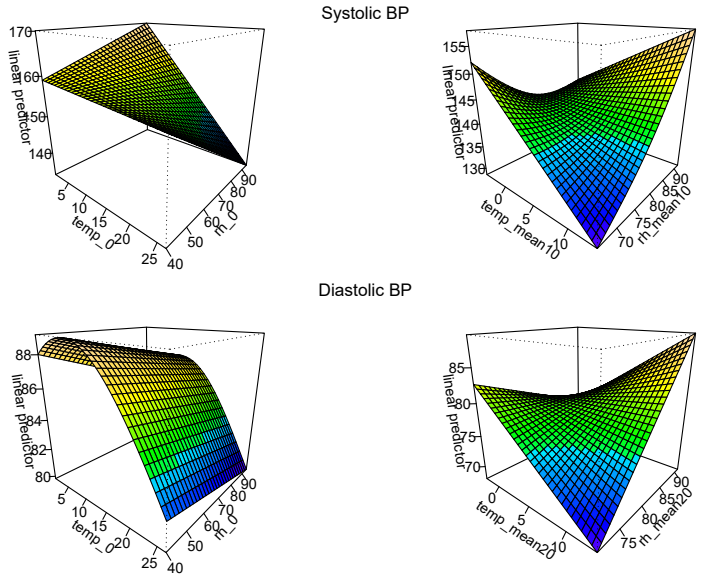

FIGURE 1. Bivariate tensor product of daily temperature and humidity. *Left panels*: Warmer months, *Right panels*: Colder months.

quantified the interaction effect between temperature and relative humidity using a bivariate tensor spline. To account for the repeated measurement structure, we incorporated the random effects. As various studies indicate the importance of seasonal variations in blood pressures, we subset our study population into warmer and colder months (see Rosenthal, T., 2004). Our data suggest that individual-specific covariates such as age and location influence blood pressures. We also see a significant effect of the interaction between climatic factors on blood pressure. Our finding of systolic and diastolic blood pressures modelled best with climatic data taken on the day of examination for warmer days and longer lags for colder days confirms earlier research studies (Brook et al., 2011). However, note that our study has a few limitations. First, the blood pressures have different scales of measurement for each follow-up. Although this seems to be a small-scale impact on the results, it may still induce some bias. Second, this study does not include the popular methods used in literature to detect interaction effects. For example, stratification of the climatic factors can give a more comprehensive and quantitative comparison of temperature and relative humidity effects on blood pressure. While GAMs are widely used in studying the effect of environmental factors on blood pressure, this study considers the interaction among these environmental factors through a bivariate tensor product spline.

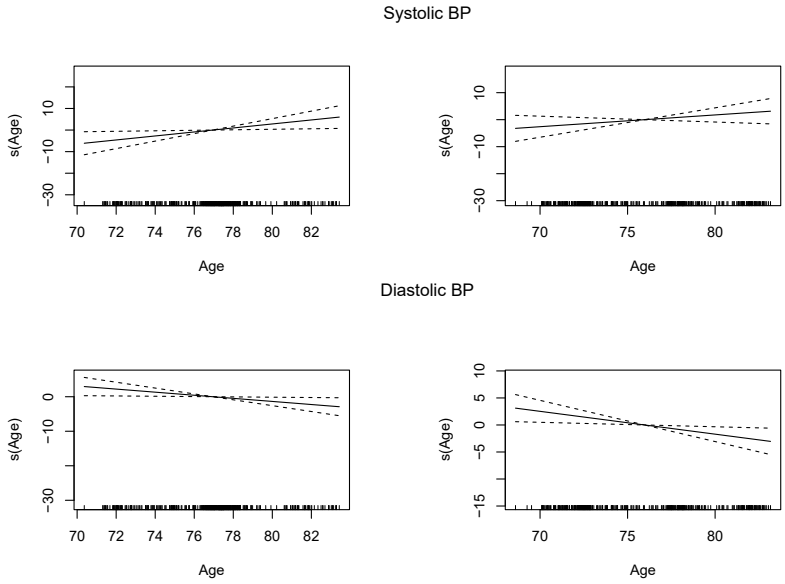

FIGURE 2. Relationship between age and blood pressures. *Left panels*: Warmer months, *Right panels*: Colder months. Estimated effects in solid lines with 95 % confidence intervals in dashed lines.

References

Barnett et al., (2007). The effect of temperature on systolic blood pressure. *Blood Pressure Monitoring*, **12**, 195 – 203.

Brook, R. D. (2017). The environment and blood pressure. *Cardiology clinics*, **35(2)**, 213-221.

Eilers, P.H and Marx, B.D (2021). *Practical Smoothing: The Joys of P-splines*. Cambridge: Cambridge University Press.

Pinto E. (2007). Blood pressure and ageing. *Postgraduate medical journal*, **83(976)**, 109–114.

Ravindra et al., (2019). Generalized additive models: Building evidence of air pollution, climate change and human health). *Environment International*, **132**, 104987.

Rosenthal, T. (2004). Seasonal variations in blood pressure. *The American journal of geriatric cardiology*, **13(5)**, 267-272.

Wood, S. N. (2006). *Generalized additive models: an introduction with R*. chapman and hall/CRC.

Article II

Complex Synergistic Effects of Air Pollution and Temperature on Blood Pressure: Evidence from the SALIA Cohort Study

Dayasri Ravi[a], Andreas Groll[a], Claudia Wigmann[b], Nidhi Singh[b], Tamara Schikowski[b,c*]

[a]*Department of Statistics, TU Dortmund University, Vogelpothsweg 87, Dortmund, 44227, NRW, Germany*
[b]*IUF-Leibniz Research Institute for Environmental Medicine, Auf 'm Hennekamp 50, Düsseldorf, 40225, NRW, Germany*
[c]*School of Public Health, Department of Environment and Health, University of Bielefeld, 33501 Bielefeld, Germany*

**Corresponding author:*
Prof. Tamara Schikowski
IUF-Leibniz Research Institute for Environmental Medicine
Email: tamara.schikowski@iuf-duesseldorf.de

ABSTRACT

Background: Studies have shown how air pollution and temperature affect blood pressure. We investigated the combined effect of air pollution and temperature on blood pressure in a cohort of older German women.

Methods: We analysed systolic (SBP) and diastolic blood pressure (DBP) data from the follow-up of the Study on the influence of Air pollution on Lung function Inflammation and Ageing (SALIA) cohort. Short-term data on air pollutants and temperature were obtained from the German Weather Bureau and the German Environment Agency. A generalized additive model was used to capture their combined effect. Stratified analyses were performed to quantify the variation in the estimated effects. The core model was adjusted for several covariates.

Results: We observed a combined effect of temperature and air pollution on blood pressure. We found that low temperatures and high levels of air pollutants such as $PM_{2.5}$ and NO_2 increased SBP and DBP. However, higher exposure to O_3 generally lowered SBP and DBP. Stratified analyses showed that temperature and air pollution had significant combined effects on women living in urban areas and with low socio-economic backgrounds.

Conclusion: The combined effect of low ambient temperature and high air pollution substantially increased BP among older German women.

Keywords: Air pollution, Temperature, Blood pressure, Simultaneous exposure, Linear and non-linear interactive effects

INTRODUCTION

Every year, an estimated seven million premature deaths are attributed to the combined effects of household and ambient air pollution [1]. There is some evidence, from studies conducted over short or long periods, that air pollution contributes to increased mortality and morbidity risks associated with hypertension, cardiovascular disease, respiratory problems and stroke, as well as frequent emergency visits for hypertension [2, 3]. Possible explanations for this association include elevated oxidative stress levels, systemic inflammation, endothelial dysfunction, alterations in blood coagulability, and the autonomic nervous system [4, 5]. Similarly, extreme temperatures have been associated with blood pressure (BP) variations leading to more cardiovascular deaths [6, 7]. A major concern is that different locations and age groups have been studied, leading to different results.

One of the reasons why such studies have failed to provide concrete certainty may be the exclusion of synergistic interactions among various environmental exposures. While the independent effects of temperature and outdoor pollutants on health are well-established, little effort has been made to explore their combined effects. The relationship between temperature and air pollution is influenced by factors such as location, climate, emissions, and specific pollutants. High temperatures can alter chemical reactions in the atmosphere, especially those involving nitrogen oxides [8, 9, 10]. On hot, sunny days, tropospheric ozone and other secondary pollutants often exceed normal levels. This combination of elevated temperatures and ozone can affect the body's thermoregulation, potentially leading to decreased BP [11]. Therefore, understanding how air quality interacts with temperature and impacts health is critical. Recent epidemiological research emphasizes the urgency of investigating the health impacts of these environmental interactions [12].

To address this research gap, we hypothesized a complex synergy between temperature and air pollution and their effects on BP in older females. Evidence suggests that older adults are more vulnerable to the harmful impacts of climate change [13]. Therefore, we utilized the well-characterized SALIA (Study on Air pollution, Lung function, Inflammation, and Aging) cohort study and conducted subgroup analyses to identify older individuals particularly susceptible to

environmental exposures. Instead of using simple multiplication as in previous studies [8], we employed a bivariate tensor product approach to capture the intricate correlations between environmental stressors. Our research utilized Generalized Additive Models (GAMs) [14] with tensor product terms to accurately describe the complex interplay between temperature, air pollutants, and BP. These flexible models can capture non-linearities and incorporate multiple factors, demonstrating superior predictive power over linear models for assessing environmental health impacts. Thus, our study is novel in its application of robust statistical techniques like GAMs to identify the combined effects of temperature and air pollution on BP.

MATERIAL AND METHODS

Study design and population

The SALIA cohort study was established between 1985 and 1994 to investigate the health effects of air pollution exposure in women by the State Government of North-Rhine Westphalia, Germany. Women from the region's industrialized Ruhr Area (Dortmund, Duisburg, Herne, Gelsenkirchen, and Essen) and two non-industrialized rural communities north of the Ruhr Area (Borken and Dülmen) make up the study population. The details of the study have been published elsewhere [15, 16]. In the present study, we utilized data from follow-up 3 (2012-2013) with 541 women. The study was approved by the ethics committee of the Medical Faculty of the Heinrich Heine University Düsseldorf (Germany; Registration number: 3507). All methods were performed in accordance with the relevant guidelines. The Declaration of Helsinki Principles was followed, and all women gave their written informed consent before the investigation.

Assessment of BP and other covariates

During the third follow-up, participants were interviewed for a thorough medical history, including diabetes mellitus, respiratory health conditions, cardiovascular diseases, medication, and lifestyle factors, following the standard study protocol. A sphygmomanometer, Omron 705 IT at the left upper arm, measured BP at a sitting position after a rest period of at least 5 minutes. The final BP value was, as a rule, defined as the mean value of second and third measurement. Systolic (SBP) and diastolic BP (DBP) were used as BP markers.

We obtained information about other potentially confounding factors such as smoking status, current passive smoking at home, the packyears (number of cigarette packs consumed per day times the duration of smoking in years), urban/rural living, age, and body mass index (BMI). The maximum length of schooling (< 10 years, 10 years, and > 10 years) attained by the participating woman or her husband was used as an indicator of socioeconomic status (SES). The participants were also classified into physically active and inactive participants according to their self-report on having ever regularly done sports. Additionally, the frequency of alcohol consumption was used as a dichotomized variable (less than once a week vs. once a week or more often). Information on diabetes was included in sensitivity analyses.

Exposure assessment

Individual exposure to short-term air pollution and temperature was estimated through average levels at women's residential addresses. In this study, we employed data from lag 0, ensuring that BP and exposures were derived from the same day without incorporating any lags.

The daily mean temperatures (Tmean) and relative humidity (RH) were obtained from COSMO-REA6 [17] at a spatial resolution of 6x6 km for the study region.

We obtained data for daily air pollution levels ($PM_{2.5}$ NO_2, and O_3) at a spatial resolution of 2x2 km based on the method of optimal interpolation from the German Environment Agency [18]. All parameters were assigned to the participant's home address.

Statistical analysis

We used a generalized additive model (GAM) with an identity link to analyze the association between different BP measurements (SBP or DBP), Tmean, and air pollutants ($PM_{2.5}$, NO_2, or O_3). The BP measurements were log-transformed to justify normality and stabilize the variance. To rule out potential confounding effects, we adjusted the models for the following covariates such as age, BMI, location (urban/rural), SES (low/medium/high), fossil heating (yes/no), packyears, current smoking and passive smoking. Additionally, all models included the season (warm: April-September, cold: October-March) as a binary temporal variable. Finally, all models contained an indicator variable for RH greater than 80%. In the final analysis, we used the complete case analysis with complete information on 541 participants.

We investigated the combined effect of Tmean and air pollutants on BP through bivariate tensor product penalized splines. Therefore, we modeled BP as a bivariate function of Tmean and one air pollutant at a time. We adjusted for the mentioned covariates and applied a smoothing effect to certain continuous variables, such as age and the packyears.

$$E[\mathrm{Y}|\mathrm{x}] = te(Tmean,\ PM_{2.5}/NO_2/O_3) + s(age) + s(packyears) + BMI + location + SES + RH_{bin} + fossil\ heating + current\ smoker + passive\ smoking + season$$

where Y represents the BP with expectation $E[\mathrm{Y}|\mathrm{x}]$ (conditional on all covariates x), $te(\cdot)$ is the bivariate tensor product of Tmean and one of the air pollutants, and $s(\cdot)$ symbolizes the (smooth) non-linear covariate effects based on penalized splines.

We performed several sensitivity analyses to test the core model's robustness. We repeated the main analysis without applying the log transformation of BP. We additionally adjusted for other covariates in the main model, like frequency of alcohol intake, physical activity, and diabetes.

Furthermore, we examined effect modifications based on stratified analyses using individual characteristics such as location, BMI, and SES.

The strength of the non-linearity of GAM estimates was tested through both Generalized cross-validation (GCV) scores and Restricted Maximum Likelihood Estimation (REML).

All analysis were performed using R statistical software, V4.3.3 [19].

RESULTS

Descriptive results

Detailed characteristics of the study population are shown in Table 1. The average age of the study participants was 77.5 (± 3.15). Most of the participants had a BMI above 25 (79.4%) (Table S4), medium to high SES (81.3%), and were non-smokers (96.9%). The proportions of participants from urban and rural areas were relatively similar.

The description of temperature, air pollutants, and BP measurements is shown in Table 2. The daily average values were as follows: Tmean was 10.4 (± 7.56)°C, $PM_{2.5}$ was 15.2 (± 9.04) µg/m^3, NO_2 was 23.5 (± 10.9) µg/m^3 and O_3 was 40.9 (± 20.5) µg/m^3. While the means across the study sample of the daily averages for all air pollutants were below or very close to the World Health Organization (WHO) Air Quality Guidelines, individual days with levels exceeding these limits

were observed. Additionally, during the time of follow-up 3, the 2005 WHO guideline for $PM_{2.5}$ was higher (25 μg/m³), and there was no established limit for NO_2.

The average SBP was generally observed to be high at around 145 (± 20.8) mmHg. However, the DBP was in the normal range (less than 80 mmHg, [20]). The average DBP was 77.8 (± 10.1) mmHg.

Figure F1 shows the individual relationships between BP and Tmean, as well as BP and air pollutants. We did not observe any general pattern among the exposures and BP. The estimated individual associations were mainly linear.

The combined effect of temperature and air pollution on BP

Figures 1A and 2A display the3D surface plots of SBP and DBP as functions of Tmean and air pollutants in the fitted model. These plots enable us to identify the relationships between Tmean and BP at any given concentration of air pollutants and between air pollutants and BP at every Tmean. Figures 1B and 2B show the associations between BP and air pollutants at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the Tmean distribution. Overall, we observed a smooth, complex, non-linear combined effect of Tmean and air pollution on BP.

The combined effect of temperature and $PM_{2.5}$ on BP: We observed higher SBP and DBP at low temperatures, which increased with increasing concentrations of $PM_{2.5}$, and a decrease in SBP at high temperatures with rising $PM_{2.5}$ levels. The association between temperature and BP was more linear for SBP and non-linear for DBP (Fig 1A, 2A). The percentile graph, similarly, shows an increase in SBP and DBP at low temperatures with an increase in exposure to $PM_{2.5}$ (Fig 1B, 2B).

The combined effect of temperature and NO_2 on BP: We observed higher SBP at low temperatures, which increased with increasing NO_2 concentrations. In contrast, the SBP decreases at higher temperatures, even with a rise in NO_2 levels (Fig 1A). However, we noticed a non-linear association between NO_2 and SBP across all temperature categories. At low and high temperatures, the association between DBP and NO_2 is linear in that DBP decreases with an increase in NO_2 (Fig 2A). In contrast, at medium temperatures, DBP increases with an increase in NO_2. The same is visible in the two-dimensional plots (Fig 1B, 2B)

The combined effect of temperature and O_3 on BP: In general, there was a negative linear association between SBP and temperature/O_3, so there was a decrease in SBP with an increase in temperature and O_3 (Fig 1A). A similar association was also observed for DBP, with a slight non-linear association with temperature (Fig 2A). Consequently, both BP levels decreased across all temperature categories, except for very low temperatures, in response to O_3 exposure. However, the slope or intensity of the decrease varied with temperature (Fig 1B, 2B).

Supplementary Table S1 represents the significance of the bivariate tensor product on BP in terms of the *p*-value of the tensor product between temperature and air pollution. We noticed that the combined effect of Tmean and air pollution is significantly associated with SBP, while it was in part insignificant with DBP. Overall, the association between air pollution and BP differs with varying levels of Tmean, indicating an interactive effect of Tmean and air pollution on BP.

Stratification analysis

We studied the combined effect of Tmean and air pollution on BP, stratified by location, SES, and BMI (Supplementary Tables S2, S3, and S4). A BMI above 25 kg/m² is classified as high by WHO guidelines. Overall, we found significant combined effects of Tmean and air pollutants on both SBP and DBP in women living in urban areas and those with low to medium SES. However, no significant combined effect was observed among older individuals with high BMI.

Sensitivity analysis

We performed several sensitivity analyses to test the robustness of the results. Controlling for additional covariates such as alcohol intake frequency, physical activity, and diabetes did not alter the findings from the core model. Additionally, we assessed the degree of non-linearity in the bivariate exposure-response surface by estimating the effective degrees of freedom (edf's) using both GCV scores and the REML method in the GAM model. Larger edf values indicate greater non-linearity, while an edf of 1 signifies a linear surface. The edf for the combined effect of Tmean and $PM_{2.5}$ concentration was estimated at 4.02 (REML) and 4.18 (GCV), respectively.

DISCUSSION

Our study within the SALIA cohort examines the non-linear relationships between temperature, air pollution, and BP. We found that the combined effects of temperature and air pollution significantly impact BP in older adults, even after controlling for confounders like individual characteristics, season, and relative humidity. To our knowledge, this is the first study to utilize a bivariate non-linear exposure-response surface to model the combined impact of air pollutants and temperature on BP. Specifically, exposure to low temperatures alongside higher concentrations of pollutants such as $PM_{2.5}$ and NO_2 was associated with increased BP.

Previous studies have examined the short-term effects of temperature and air pollution on BP, often assuming linear relationships, with inconsistent findings: some report positive association [2, 21, 22], others negative [23, 24]. However, most of these studies have only examined the individual impacts of temperature and pollution on BP, with limited investigation of their combined effects. While some studies incorporated linear interaction terms between temperature and pollutants, indicating that low temperatures and high pollution may elevate BP in healthy adults [25], this approach struggles to capture non-linearity and often yields non-significant results. A few studies investigating short- [26] and long-term [27] exposures, have identified non-linear associations between certain pollutants (O_3, NO_2) and SBP, but not with particulate matter. Thus, to address non-linearity, our study took a more comprehensive approach by examining the impact of $PM_{2.5}$, NO_2, and O_3 exposures on BP, modified by temperature, while considering the possibility of both linear and non-linear associations. Consistent with the previous studies, our study also indicated a linear combined effect of temperature and $PM_{2.5}$ on SBP, while the combined effect with NO_2 was non-linear with BP.

We observed an elevation in BP for exposures of $PM_{2.5}$ within the low-temperature levels and an opposite effect of a decrease in BP for exposures of $PM_{2.5}$ within the medium- and high-temperature levels. While most of the previous studies report an elevation in BP for exposure to $PM_{2.5}$ [28,21] and a few others [29] reported no significant impact of $PM_{2.5}$ in BP, we noticed the substantial role of the temperature modification. The relationship between BP and other air pollutants is also quite inconsistent in previous studies. A Danish cohort study and others have reported decreased SBP with exposure to NO_2 [30, 31]. Although we observed decreased DBP at low and high temperatures, DBP increased with NO_2 exposure at medium temperatures, while the association with SBP was non-linear. Despite observing a generally negative association between O_3 and BP, our findings contrast with studies showing increased DBP from O_3 exposure [32, 33]. These inconsistencies may be due to differences in populations, regions, sample sizes, and the lack of consideration for temperature's modifying effect. Another reason for these discrepancies is the limitation of not accounting for non-linear relationships, as we observed that the association between air pollution and BP was partially non-linear and varied with temperature.

Our analysis revealed that temperature and air pollution had a stronger impact on women in urban areas compared to rural ones. Previous research also reported increased BP with moves from suburban to metropolitan areas, likely due to higher carbonaceous $PM_{2.5}$ levels in urban

environments [34]. Our findings of significant associations for low SES groups align with studies that connect air pollution to hypertension among women in lower-income, less-educated, and rural populations [35].

Much remains to be understood about the biological processes connecting air pollution and BP. One widely accepted theory suggests that PM induces oxidative reactions and systemic inflammation, leading to vascular dysfunction [36]. Aging arteries have reduced antioxidant capacity, resulting in heightened oxidative stress. This compromises endothelial and vasomotor functions, impairing BP autoregulation during inflammation [37]. This may explain why older adults experience greater BP increases from air pollution than younger individuals [38]. Ultrafine particles may also penetrate alveolar walls, directly affecting endothelial cells and raising BP [39]. Similarly, the mechanisms linking temperature and BP are unclear. Cold temperatures stimulate the nervous system, increasing heart rate and BP, whereas higher temperatures may lower BP, possibly due to reduced peripheral resistance [7].

Our study has several strengths. Firstly, modeling a two-dimensional exposure-response surface between temperature and air pollution on BP is statistically robust, as it accommodates differing units of environmental variables and captures both individual and interactive effects. While other linear models may yield similar insights, bivariate tensor products reduce power loss from temperature stratification [40] and support non-linear or complex relationships. Secondly, our well-characterized cohort, including factors like smoking and drinking, provides a solid basis for controlling confounding variables.

We also acknowledge several limitations of our study. The study population consisted of older women from the industrialized area of Germany. Thus, the conclusions may not be generalized to other study populations. Secondly, we could not rule out the possibility of unmeasured confounding variables, such as psychosocial stressors, which might impact BP. Thirdly, since the participants were primarily more than 60 years of age, controlling for antihypertensive medication could have been desirable.

CONCLUSION

In conclusion, we found an increase in SBP and DBP with exposure to low temperature and higher levels of air pollution in older German women. Unlike most previous studies, which use the linear interaction model between temperature and air pollution, we propose using non-linear bivariate tensor splines to model their combined effect. Future studies should explore the combined effects of temperature and air pollution on BP and investigate biological pathways to clarify these complex interactions. The findings of this research highlight the urgent need to address the intricate synergy between environmental variables and public health outcomes.

REFERENCES

[1] World Health Organization. Air pollution — who.int. *https://www.who.int/news-room/fact-sheets/detail/ambient-(outdoor)-air-quality-and-health* (2022). Accessed 8 Feb. 2023.

[2] Brook, R. D. & Rajagopalan, S. Particulate matter, air pollution, and blood pressure. *J. Am. Soc. Hypertens.* **3,** 332–350 (2009).

[3] Yang, B. Y. et al. Global association between ambient air pollution and blood pressure: a systematic review and meta-analysis. *Environ. Pollut.* **235,** 576–588 (2018).

[4] Nemmar, A., Subramaniyan, D., Yasin, J. & Ali, B. H. Impact of experimental type 1 diabetes mellitus on systemic and coagulation vulnerability in mice acutely exposed to diesel exhaust particles. *Part. Fibre Toxicol.* **10,** 1–10 (2013).

[5] Wilson, S. J., Miller, M. R. & Newby, D. E. Effects of diesel exhaust on cardiovascular function and oxidative stress. *Antioxid. Redox Signal.* **28,** 819–836 (2018).

[6] Modesti, P. A. Season, temperature and blood pressure: a complex interaction. *Eur. J. Intern. Med.* **24,** 604–607 (2013).
[7] Wang, Q. et al. Environmental ambient temperature and blood pressure in adults: a systematic review and meta-analysis. *Sci. Total Environ.* **575,** 276–286 (2017).
[8] Analitis, A. et al. Synergistic effects of ambient temperature and air pollution on health in Europe: results from the PHASE project. *Int. J. Environ. Res. Public Health* **15,** 1856 (2018).
[9] Mokoena, K. K. et al. Interaction effects of air pollution and climatic factors on circulatory and respiratory mortality in Xi'an, China between 2014 and 2016. *Int. J. Environ. Res. Public Health* **17,** 9027 (2020).
[10] Zhou, L. et al. The interactive effects of extreme temperatures and PM2.5 pollution on mortalities in Jiangsu Province, China. *Sci. Rep.* **13,** 9479 (2023).
[11] De Vita, A. et al. The impact of climate change and extreme weather conditions on cardiovascular health and acute cardiovascular diseases. *J. Clin. Med.* **13,** 759 (2024).
[12] Areal, A. T., Zhao, Q., Wigmann, C., Schneider, A. & Schikowski, T. The effect of air pollution when modified by temperature on respiratory health outcomes: a systematic review and meta-analysis. *Sci. Total Environ.* **811,** 152336 (2022).
[13] Simoni, M. et al. Adverse effects of outdoor pollution in the elderly. *J. Thorac. Dis.* **7,** 34 (2015).
[14] Wood, S. N. *Generalized additive models: an introduction with R.* (Chapman and Hall/CRC, 2017).
[15] Schikowski, T. et al. Long-term air pollution exposure and living close to busy roads are associated with COPD in women. *Respir. Res.* **6,** 1–10 (2005).
[16] Ohlwein, S. et al. Air pollution and diastolic function in elderly women – Results from the SALIA study cohort. *Int. J. Hyg. Environ. Health* **219,** 356–363 (2016).
[17] Bollmeyer, C. et al. Towards a high-resolution regional reanalysis for the European CORDEX domain. *Q. J. R. Meteorol. Soc.* **141,** 1–5 (2015).
[18] Minkos, A., Dauert, U., Feigenspan, S. & Kessenger, S. Air Quality 2016: Preliminary Evaluation. Dessau-Roßlau: Umweltbundesamt; (2017).
[19] R Core Team. R: A Language and Environment for Statistical Computing. *R Foundation for Statistical Computing.* https://www.R-project.org/ (2024).
[20] Chobanian, A. V. et al. Seventh report of the joint national committee on prevention, detection, evaluation, and treatment of high blood pressure. *Hypertension* **42,** 1206–1252 (2003).
[21] Ishii, M. et al. Association of short-term exposure to Asian dust with increased blood pressure. *Sci. Rep.* **10,** 17630 (2020).
[22] Wen, T. et al. Short-term air pollution levels and blood pressure in older women. *Epidemiology* **34,** 271–281 (2023).
[23] Chen, Q. et al. Association between ambient temperature and blood pressure and blood pressure regulators: 1831 hypertensive patients followed up for three years. *PLoS One* **8,** e84522 (2013).
[24] Giorgini, P. et al. Particulate matter air pollution and ambient temperature: opposing effects on blood pressure in high-risk cardiac patients. *J. Hypertens.* **33,** 2032–2038 (2015).
[25] Wu, S. et al. Does ambient temperature interact with air pollution to alter blood pressure? A repeated-measure study in healthy adults. *J. Hypertens.* **33,** 2414–2421 (2015).
[26] Choi, Y. J. et al. Short-term effects of air pollution on blood pressure. *Sci. Rep.* **9,** 20298 (2019).
[27] Arku, R. E. et al. Long-term exposure to outdoor and household air pollution and blood pressure in the Prospective Urban and Rural Epidemiological (PURE) study. *Environ. Pollut.* **262,** 114197 (2020).
[28] Brook, R. D. et al. Differences in blood pressure and vascular responses associated with ambient fine particulate matter exposures measured at the personal versus community level. *Occup. Environ. Med.* **68,** 224–230 (2011).

[29] Lee, D. H. et al. Personal exposure to fine particulate air pollutants impacts blood pressure and heart rate variability. *Sci. Rep.* **10,** 16538 (2020).
[30] Sørensen, M. et al. Long-term exposure to traffic-related air pollution associated with blood pressure and self-reported hypertension in a Danish cohort. *Environ. Health Perspect.* **120,** 418–424 (2012).
[31] Floyd, C. N. et al. Acute blood pressure-lowering effects of nitrogen dioxide exposure from domestic gas cooking via elevation of plasma nitrite concentration in healthy individuals. *Circ. Res.* **127,** 847–848 (2020).
[32] Chuang, K. J., Yan, Y. H. & Cheng, T. J. Effect of air pollution on blood pressure, blood lipids, and blood sugar: a population-based approach. *J. Occup. Environ. Med.* **52,** 258–262 (2010).
[33] Song, J. et al. Short-time exposure to ambient ozone and associated cardiovascular effects: a panel study of healthy young adults. *Environ. Int.* **137,** 105579 (2020).
[34] Wu, S. et al. Blood pressure changes and chemical constituents of particulate air pollution: results from the healthy volunteer natural relocation (HVNR) study. *Environ. Health Perspect.* **121,** 66–74 (2013).
[35] Abba, M. S. et al. Household air pollution and high blood pressure: a secondary analysis of the 2016 Albania Demographic Health and Survey dataset. *Int. J. Environ. Res. Public Health* **19,** 2611 (2022).
[36] Arias-Pérez, R. D. et al. Inflammatory effects of particulate matter air pollution. *Environ. Sci. Pollut. Res.* **27,** 42390–42404 (2020).
[37] Zhou, X., Bohlen, H. G., Unthank, J. L. & Miller, S. J. Abnormal nitric oxide production in aged rat mesenteric arteries is mediated by NAD(P)H oxidase-derived peroxide. *Am. J. Physiol. Heart Circ. Physiol.* **297,** H2227–H2233 (2009).
[38] Nicolaou, L. et al. Cross-sectional analysis of the association between personal exposure to household air pollution and blood pressure in adult women: Evidence from the multi-country Household Air Pollution Intervention Network (HAPIN) trial. *Environ. Res.* **214,** 114121 (2022).
[39] Schraufnagel, D. E. The health effects of ultrafine particles. *Exp. Mol. Med.* **52,** 311–317 (2020).
[40] Royston, P., Altman, D. G. & Sauerbrei, W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat. Med.* **25,** 127–141 (2006).

ACKNOWLEDGMENTS

This work has been supported by the Research Training Group “Biostatistical Methods for High-Dimensional Data in Toxicology” (RTG 2624, Project R2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—Project Number 427806116). The SALIA cohort study was supported by the Deutsche Forschungsgemeinschaft (DFG; HE-4510/2-1, KR, 1938/3-1, LU 691/4-1 and SCHI 1358/3-1), the Ministry of the Environment of the state North Rhine-Westphalia (Düsseldorf, Germany), the Federal Ministry of the Environment (Berlin, Germany), the German Statutory Accident Insurance (DGUV) under Grant No: 617.0-FP266, the European Community's Seventh Framework Program (FP7/2007–2011) under grant agreement number 211250, the German Federal Ministry of Education and Research (BMBF) and by the Research Commission of the Medical Faculty of the Heinrich Heine University of Düsseldorf (FoKo 9772465).
The IUF is funded by the federal and state governments - the Ministry of Culture and Science of North Rhine-Westphalia (MKW) and the Federal Ministry of Education and Research (BMBF).

AUTHOR CONTRIBUTIONS

DR wrote the main manuscript and contributed to the overall structure. DR, AG and CW performed data analysis and statistical modeling. NS and TS ensured the interpretation of results. All authors contributed to manuscript revisions, reviewed, and approved the final version of the manuscript.

DATA AVAILABILITY

The datasets used and/or analysed during the current study available from the corresponding author on reasonable request.

COMPETING INTERESTS

The authors declare no competing interests.

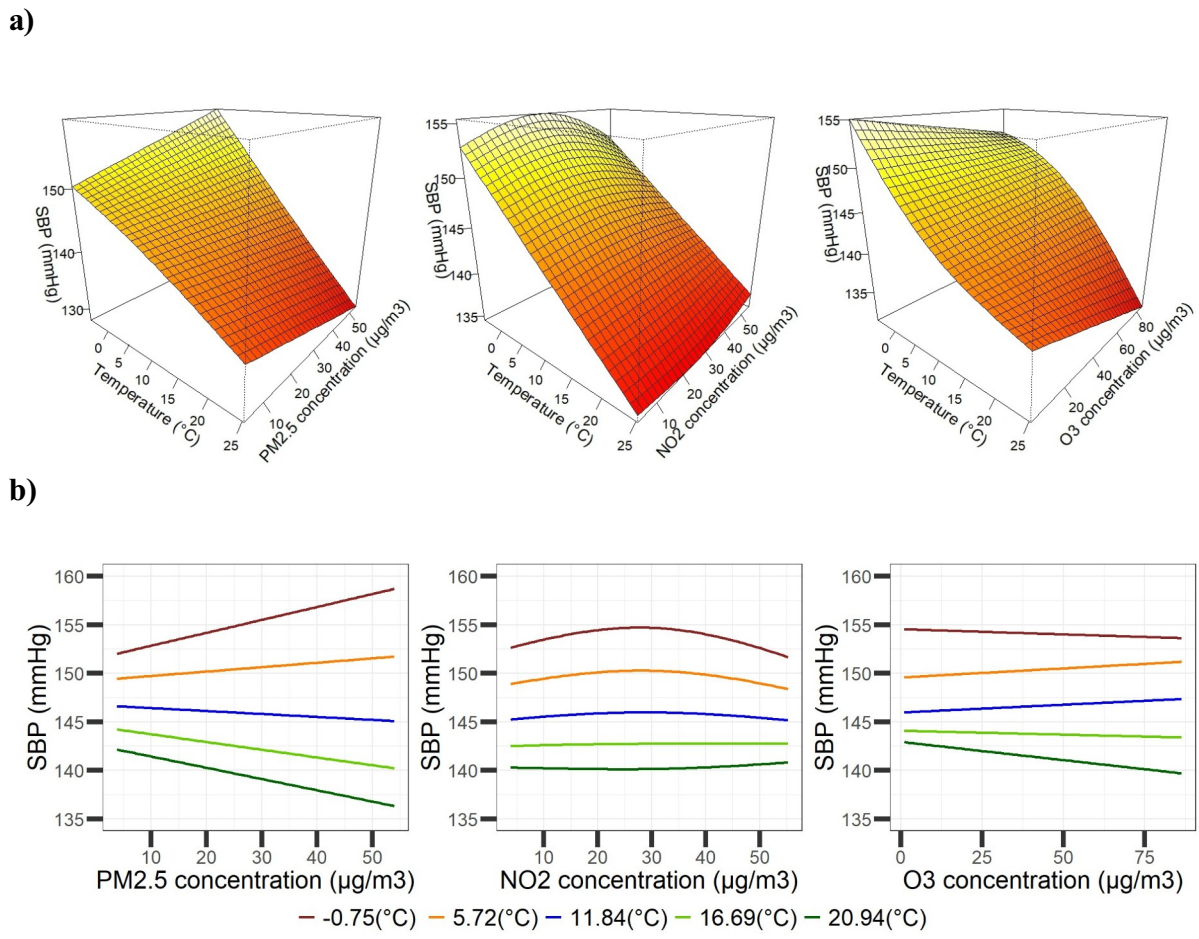

Figure 1: (*Panel* **a**) Visual representation of the bivariate response surfaces for SBP of temperature and air pollutants, $PM_{2.5}$, NO_2, and O_3 (*column-wise*), estimated by the tensor product model. All models were adjusted for age, body mass index (BMI), socioeconomic status (SES), location (urban/rural), packyears, current smoking (yes/no), second smoker (yes/no), fossil fuel heating (yes/no), the season, and relative humidity (binary with cutoff 80%). The color intensity represents the magnitude of the BP values. Darker colors indicate lower values; lighter colors

represent larger values. (*Panel* **b**) The adjusted associations between SBP and air pollutants at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the Tmean distribution.

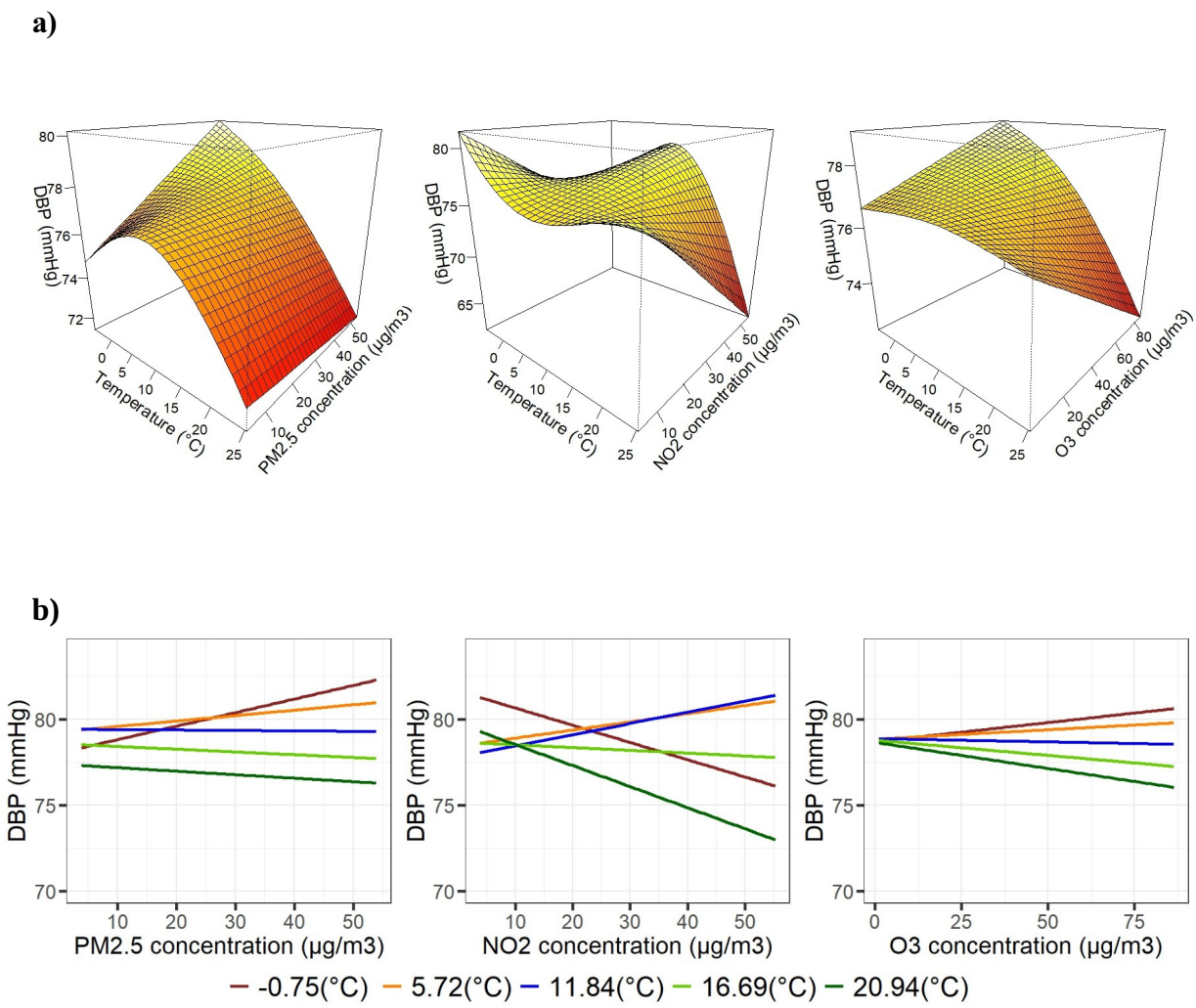

Figure 2: (*Panel* **a**) Visual representation of the bivariate response surfaces for DBP of temperature and air pollutants, $PM_{2.5}$, NO_2, and O_3 (*column-wise*), estimated by the tensor product model. All models were adjusted for age, body mass index (BMI), socioeconomic status (SES), location (urban/rural), packyears, current smoking (yes/no), second smoker (yes/no), fossil fuel heating (yes/no), the season, and relative humidity. The color intensity represents the magnitude of the BP values. Darker colors indicate lower values; lighter colors represent larger values. (*Panel* **b**) The adjusted associations between DBP and air pollutants at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, and 90^{th} percentiles of the Tmean distribution.

Table 1: Descriptive statistics of the study population.

	Overall (N=541)
Age in years	
Mean (SD)	77.5 (3.15)
Median [Min, Max]	77.5 [70.1, 83.4]
BMI in kg/m^2	
Mean (SD)	28.4 (4.52)
Median [Min, Max]	27.8 [15.7, 46.2]
Packyears [packs/day × years]	
Mean (SD)	3.44 (11.3)
Median [Min, Max]	0 [0, 133]
SES	
Low	101 (18.7%)
Medium	251 (46.4%)
High	189 (34.9%)
Location	
Rural	271 (50.1%)
Urban	270 (49.9%)
Season	
Winter	252 (46.6%)
Summer	289 (53.4%)
Diabetes	
No	465 (86.0%)
Yes	73 (13.5%)
Heating with fossil fuels	
No	482 (89.1%)
Yes	53 (9.8%)
Smoker	
No	524 (96.9%)
Yes	17 (3.1%)
Second smoker	
No	472 (87.2%)
Yes	65 (12.0%)
Regular sports	
Active	226 (41.8%)
Inactive	313 (57.9%)
Alcohol consumption	
Once a week or more often	137 (25.3%)
Less than once a week	402 (74.3%)

Table 2: Descriptive statistics of meteorological data, air pollution data, and blood pressure.

	Overall (N=541)
Temperature in °C	
Mean (SD)	10.4 (7.56)
Median [Min, Max]	10.9 [-4.52, 26.5]
Relative humidity in %	
Mean (SD)	78.0 (10.9)
Median [Min, Max]	80.5 [47.5, 95.9]
$PM_{2.5}$ in $\mu g/m^3$	
Mean (SD)	15.2 (9.04)
Median [Min, Max]	12.0 [3.90, 53.9]
NO_2 in $\mu g/m^3$	
Mean (SD)	23.5 (10.9)
Median [Min, Max]	23.6 [3.85, 55.3]
O_3 in $\mu g/m^3$	
Mean (SD)	40.9 (20.5)
Median [Min, Max]	41.4 [1.10, 86.3]
Systolic BP in mmHg	
Mean (SD)	145 (20.8)
Median [Min, Max]	144 [87.5, 248]
Diastolic BP in mmHg	
Mean (SD)	77.8 (10.1)
Median [Min, Max]	77.0 [50.0, 112]

Supplementary Material: Complex Synergistic Effects of Air Pollution and Temperature on Blood Pressure: Evidence from the SALIA Cohort Study

Dayasri Ravi[a], Andreas Groll[a], Claudia Wigmann[b], Nidhi Singh[b], Tamara Schikowski[b,c]

[a]*Department of Statistics, TU Dortmund University, Vogelpothsweg 87, Dortmund, 44227, NRW, Germany*
[b]*IUF-Leibniz Research Institute for Environmental Medicine, Auf 'm Hennekamp 50, Düsseldorf, 40225, NRW, Germany*
[c]*School of Public Health, Department of Environment and Health, University of Bielefeld, 33501 Bielefeld, Germany*

Table S1: *p*-values of the bivariate tensor product between air pollution and temperature (*p*-values<0.05 in bold).

	$PM_{2.5}$	NO_2	O_3
SBP	**0.03**	**0.04**	**0.03**
DBP	0.23	**0.04**	0.27

Table S2: Statistical significance of the combined effect between air pollution and temperature in different locations (*p*-values<0.05 in bold).

Location	BP	$PM_{2.5}$	NO_2	O_3
Rural	SBP	0.32	0.47	0.22
(N=271)	DBP	0.67	0.59	0.98
Urban	SBP	**0.03**	**0.04**	**0.01**
(N=270)	DBP	**0.02**	**0.03**	**0.02**

Table S3: Statistical significance of the combined effect between air pollution and temperature in different SES status (*p*-values<0.05 in bold).

SES	BP	$PM_{2.5}$	NO_2	O_3
Low	SBP	**0.01**	**0.00**	**0.00**
(N=101)	DBP	0.13	0.16	0.06
Medium	SBP	0.08	**0.00**	**0.01**
(N=251)	DBP	0.09	**0.02**	**0.04**
High	SBP	0.28	0.78	0.85
(N=189)	DBP	0.06	0.72	0.93

Table S4: Statistical significance of the combined effect between air pollution and temperature in different groups of BMI (p-values<0.05 in bold).

BMI	BP	$PM_{2.5}$	NO_2	O_3
Low BMI (N=111)	SBP	0.13	0.16	0.19
	DBP	0.65	0.77	0.75
High BMI (N=430)	SBP	0.15	0.14	0.18
	DBP	0.17	**0.02**	0.07

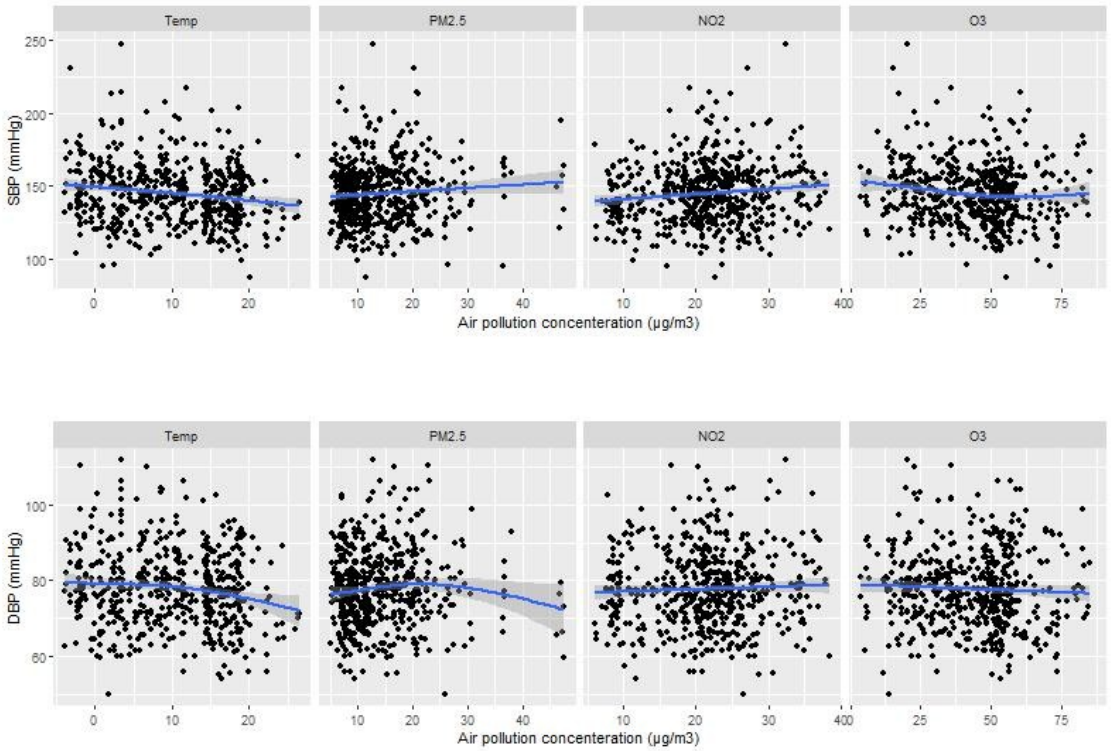

Figure F1: Individual relationships between Tmean ($^{\circ}$C) / air pollution concentrations ($\mu g/m^3$) and SBP (mmHg) / DBP (mmHg). The blue curve in the plot represents the modeled trend derived from a Generalized Additive Model (GAM) fit. For illustration, a simple GAM model with exposures and BP is used. *Shadows* represent the 95% confidence intervals (CIs).

-

Article III

Optimizing Variable Selection in Multi-Omics Datasets: A Focus on Exclusive Lasso

Dayasri Ravi( ) and Andreas Groll

Department of Statistics, TU Dortmund University, Vogelpothsweg 78, 44227 Dortmund, Germany
{ravi,groll}@statistik.tu-dortmund.de

Abstract. Multi-omics datasets pose significant challenges due to their structured nature, where highly correlated variables are grouped within a complex, high-dimensional framework. Traditional Lasso methods encounter limitations in handling correlated features within these groups effectively. To address this issue, we propose using Exclusive Lasso, focusing on inducing sparsity at the intra-group level. Additionally, we introduce an efficient algorithm for solving the related optimization problem. By prioritizing feature selection robustness within correlated group structures, our proposed methodology offers a promising solution to the challenges inherent in analyzing biological datasets. This advancement enhances our ability to extract meaningful insights from multi-omics data, thus facilitating deeper understanding and exploration of complex biological systems.

Keywords: Variable selection · Composite penalty · Exclusive Lasso

1 Introduction

Effectively utilizing multi-omics data, encompassing genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics, presents a considerable challenge. The critical characteristic of multi-omics data lies in the extensive dimensionality of the datasets. This poses specific requirements for constructing prediction models, as they must handle datasets where the number of variables significantly surpasses the number of observations. Another challenge in working with multi-omics data is its structured nature, where variables are grouped into non overlapping categories. This structural organization should be considered when constructing prediction models. A crucial aspect of multi-omics data is the high dimensionality within each omic group of datasets. In such scenarios, incorporating the group structure during model building or prioritizing clinical variables becomes beneficial. Otherwise, clinical variables which are very few in number may be overshadowed by the vast amount of omics data. It is widely recognized that a substantial portion of omics data lacks informativeness for prediction due to redundancy or irrelevance [1,2]. The importance of feature

J. Einbeck et al. (Eds.): IWSM 2024, CONTRIB.STAT., pp. 142–147, 2024.
https://doi.org/10.1007/978-3-031-65723-8_22

selection is indisputable, often requiring sparse and interpretable models that include only a limited number of variables.

Various group regularizers, including the well-known Group Lasso [3], have been extensively explored to address this challenge. Group Lasso employs $L_{2,1}$-norm regularization to promote sparsity between groups. This means that features within the same group, such as genomics or proteomics, are either selected or excluded together. However, genomics data are typically grouped together due to their high correlation. Incorporating inter-group sparsity could lead to overlooking this interdependence among groups. We propose the Exclusive Lasso [4], incorporating an $L_{1,2}$-norm regularization term to address this challenge. This term promotes sparsity within groups while allowing for a relaxation of sparsity between groups, ensuring that at least one variable from each group is selected. We introduce the *Newton Method-$L_{1,2}$ Sparsity Algorithm* (*NM-$L_{1,2}$ Sparsity Algorithm*) as an alternative estimation approach for the Exclusive Lasso, employing a quadratic approximation of the $L_{1,2}$-norm.

The paper is organized as follows. Section 2 introduces the Exclusive Lasso technique and the new estimation procedure. Results from simulation studies and real-data applications are presented in Sect. 3 and Sect. 4 concludes.

2 Methods

For $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$, we consider the problem of the form

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = L(\boldsymbol{\beta}) + \frac{\lambda}{2} \sum_{g \subset \mathcal{G}} \left(\sum_{k \in g} |\beta_k| \right)^2, \tag{1}$$

where $L(\boldsymbol{\beta})$ is a loss function, $\lambda \geq 0$ is the penalty parameter and $\mathcal{G}$ is a collection of non-overlapping predefined groups of indices g such that $\bigcup_{g \subset \mathcal{G}} = \{1, \ldots, p\}$. Following [5], we use a quadratic approximation for the transformation of the L_1 norm into a differentiable function given by $|x| \approx \sqrt{x^2 + c}$. Here, c is a small positive number. With this approximation, the penalty term from Eq. (1) can be rewritten as

$$P(\boldsymbol{\beta}) = \frac{1}{2} \sum_{g \subset \mathcal{G}} \left(\sum_{k \in g} \sqrt{\beta_k^2 + c} \right)^2. \tag{2}$$

Let $S = \{1, \ldots, p\}$ be the set of indices corresponding to the entries of $\boldsymbol{\beta}$. We construct square matrices $\mathbf{M}_g$ with dimensions $|S \cap g|$ such that

$$\mathbf{M}_{g_{ij}} = \begin{cases} 1, & \text{if } i = 1, \\ 0, & \text{if } i > 1. \end{cases}$$

Let $\mathbf{B} \in \mathbb{R}^{p \times 1}$ represent the vector $\left(\sqrt{(\beta_1^2 + c)}, \ldots, \sqrt{(\beta_p^2 + c)} \right)^\top$, then

$$\frac{\partial \mathbf{B}}{\partial \beta_k} = \frac{\beta_k}{\sqrt{(\beta_k^2 + c)}} \quad \& \quad \frac{\partial^2 \mathbf{B}}{\partial \beta_k^2} = \frac{c}{(\beta_k^2 + c)^{\frac{3}{2}}}.$$

We rewrite the penalty term as $0.5 \cdot (\mathbf{GB})^{\mathbf{T}}\mathbf{GB}$, where $\mathbf{G} \in \mathbb{R}^{p \times p}$ is a block diagonal matrix with matrices $\mathbf{M}_g$. Using matrix and vector differentiation rules, the gradient and Hessian of the penalty term $P(\boldsymbol{\beta})$ can be derived as

$$\frac{\partial}{\partial \beta_k} P(\boldsymbol{\beta}) = \mathbf{B}^{\mathbf{T}}(\mathbf{C} + \mathbf{C}^{\mathbf{T}}) \frac{\partial \mathbf{B}}{\partial \beta_k}$$

where $\mathbf{C} \in \mathbb{R}^{p \times p} = \mathbf{G}^{\mathbf{T}}\mathbf{G}$. With $\mathbf{D} \in \mathbb{R}^{p \times p} = \mathbf{C} + \mathbf{C}^{\mathbf{T}}$, we obtain

$$\frac{\partial^2}{\partial \beta_k^2} P(\boldsymbol{\beta}) = \left[\frac{\partial \mathbf{B}}{\partial \beta_k} \mathbf{D}^T \frac{\partial \mathbf{B}}{\partial \beta_k} + \mathbf{B}^T \mathbf{D} \frac{\partial^2 \mathbf{B}}{\partial \beta_k^2} \right].$$

We perform Newton iterations of the form $\boldsymbol{\beta}^{(l+1)} := \boldsymbol{\beta}^{(l)} - t(\nabla^2 g(\boldsymbol{\beta}^{(l)}))^{-1} \nabla g(\boldsymbol{\beta}^{(l)})$, where $g(\boldsymbol{\beta})$ is a continuous and twice-differentiable approximation function of $f(\boldsymbol{\beta})$. Following [6], we choose step size t using a back-tracking line search satisfying the Armijo condition.

3 Results

We evaluate our algorithm by comparing it with the standard Lasso [7] using simulated data and a real-world multi-omics dataset.

3.1 Simulation Studies

Following [4], we simulate two examples with $n = 100$ observations and $p = 100$ variables. In the first example, the variables are divided into four equal groups, with the true parameter being non-zero in the first index of each group. The data is simulated from a multivariate normal distribution with a Toeplitz covariance matrix with entries $\Sigma_{i,j} = a^{|i-j|}$ for variables in the same group, and $\Sigma_{i,j} = b^{|i-j|}$ for variables in different groups. For the first covariance matrix, we opt for a strong correlation of $a = b = 0.9$, simulating high correlation both within and between groups. For the second scenario, we use a moderate correlation of $a = 0.9$ and $b = 0.6$, resulting in high correlation within and medium correlation between group. Finally, for the third scenario, both within- and between-group correlations are set at a moderate level of $a = b = 0.6$. We repeat the same set of scenarios for the second example. Here, the variables are divided into the same four equal groups, with the true parameter being non-zero in more than one index of each group. In all the above scenarios, the data is simulated using the model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)$ and $\lambda = \max_i |\boldsymbol{X}_i^T \boldsymbol{y}|$. We report the results in terms of average variable selection accuracy, standard error (SE), and root mean square error (RMSE) on external test data for 50 simulations in Table 1 and 2 (Figs. 1 and 2).

Table 1. Example 1: Comparison of variable selection methods

		NM-$L_{1,2}$	Lasso
Scenario 1	Accuracy (SE)	**0.74** (0.015)	0.31 (0.015)
	RMSE	**1.35**	2.349
Scenario 2	Accuracy (SE)	**0.86** (0.022)	0.33 (0.017)
	RMSE	**1.37**	2.178
Scenario 3	Accuracy (SE)	**0.95** (0.015)	0.36 (0.018)
	RMSE	**1.65**	2.208

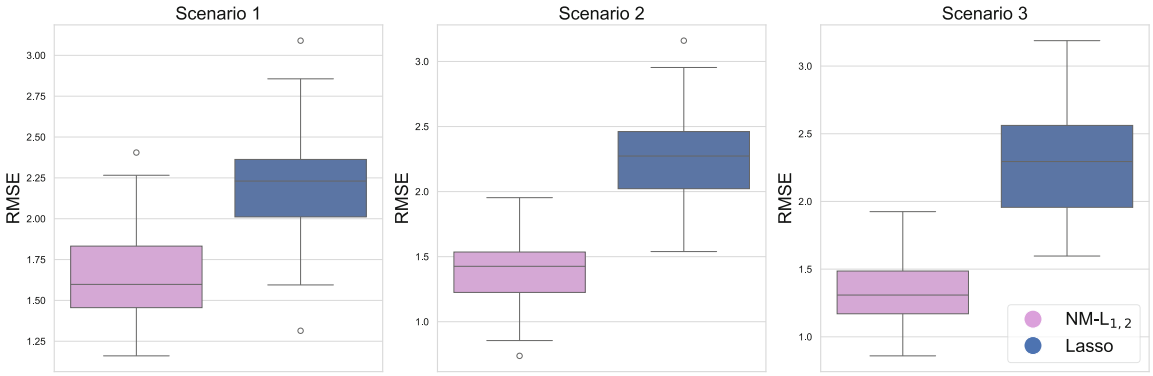

Fig. 1. Example 1: Root Mean Square Error (RMSE) for Two Models

Table 2. Example 2: Comparison of variable selection methods

		NM-$L_{1,2}$	Lasso
Scenario 1	Accuracy (SE)	**0.84** (0.024)	0.29 (0.015)
	RMSE	**2.07**	4.15
Scenario 2	Accuracy (SE)	**0.91** (0.017)	0.31 (0.015)
	RMSE	**2.22**	3.89
Scenario 3	Accuracy (SE)	**0.95** (0.011)	0.30 (0.014)
	RMSE	**2.52**	3.66

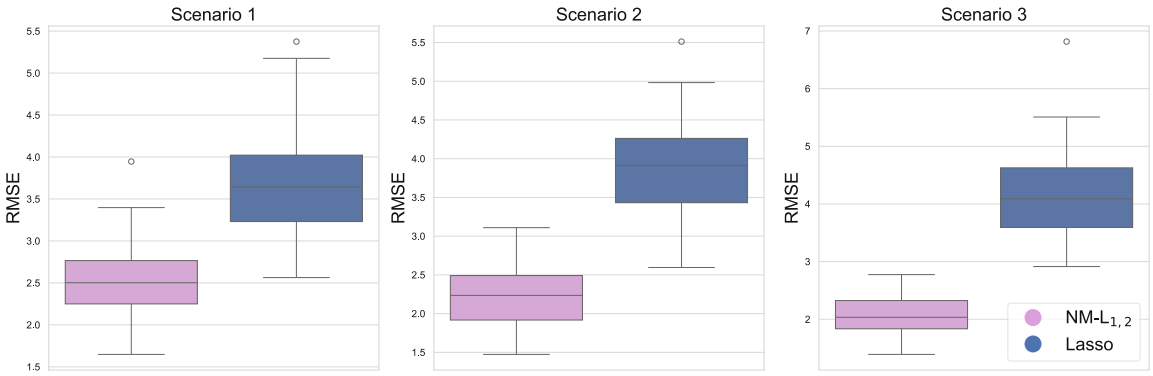

Fig. 2. Example 2: Root Mean Square Error (RMSE) for Two Models

3.2 Down Syndrome Analysis

We validate the effectiveness of the proposed variable selection algorithm using multi-omics data to predict Down Syndrome. Down Syndrome (DS) is attributed to the presence of either a complete or partial extra copy of human chromosome 21 (HSA21) [8]. It is the most common genetic condition in humans, occurring in about 1 in 700 live births. Our data comprises of 29 individuals with DS, along with their mothers and unaffected siblings. The family-centric approach is preferred as it allows for a comprehensive examination of potential influences on DNA methylation patterns, encompassing both genetic and environmental (lifestyle) factors within familial contexts. There are two molecular data types (methylation and glycomics data) and clinical data, i.e., three groups of variables. We compare *NM-*$L_{1,2}$ with Lasso and conventional Exclusive Lasso with coordinate descent from the `Exclusive Lasso` R package. We choose $\lambda = \max_i |\boldsymbol{X}_i^T \boldsymbol{y}|$ such that it is large enough to estimate one variable in each group. We randomly split the dataset into a training and a testing set, with 20% of the data for testing. Figure 3 shows the Receiver Operating Characteristic (ROC) curve for the three models. In contrast to the other two models, our approach showcases superior predictive performance, emphasizing the utilization of optimally chosen features.

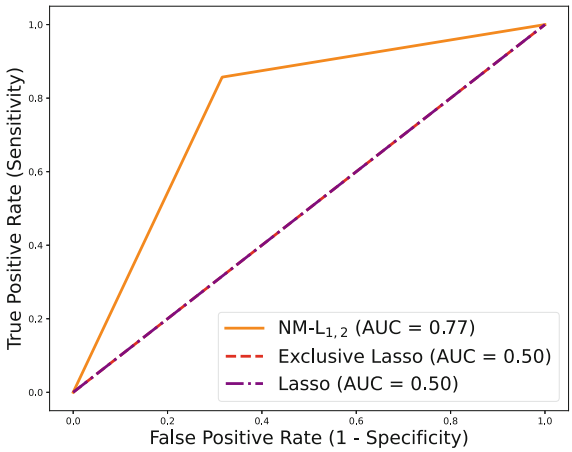

Fig. 3. Receiver Operating Characteristic (ROC) Curve for Three Models

4 Conclusion

In this work, we develop a novel algorithm to solve the Exclusive Lasso problem. Our findings demonstrate superior performance, surpassing that of standard Lasso, particularly in scenarios with highly correlated features. We propose the use of Exclusive Lasso for within-group variable selection, frequently encountered in biological datasets, especially in the context of multi-omics studies.

Acknowledgments. This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project R2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation-Project Number 427806116).

References

1. Libralon, G.L., de Carvalho, A.C.P.D.L.F., Lorena, A.C.: Pre-processing for noise detection in gene expression classification data. J. Braz. Comput. Soc. **15**, 3–11 (2009)
2. Li, Y., Mansmann, U., Du, S., Hornung, R.: Benchmark study of feature selection strategies for multi-omics data. BMC Bioinf. **23**(1), 412 (2022)
3. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat Methodol. **68**(1), 49–67 (2006)
4. Campbell, F., Allen, G.I.: Within group variable selection through the exclusive lasso (2017)
5. Oelker, M.R., Tutz, G.: A uniform framework for the combination of penalties in generalized structured models. Adv. Data Anal. Classif. **11**(1), 97–120 (2017)
6. Schmidt, M., Fung, G., Rosales, R.: Fast optimization methods for L1 regularization: a comparative study and two new approaches. In: Kok, J.N., Koronacki, J., Mantaras, R.L., Matwin, S., Mladenič, D., Skowron, A. (eds.) ECML 2007. LNCS (LNAI), vol. 4701, pp. 286–297. Springer, Heidelberg (2007). https://doi.org/10.1007/978-3-540-74958-5_28
7. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Stat. Soc. Ser. B Stat Methodol. **58**(1), 267–288 (1996)
8. Bacalini, M.G., et al.: Identification of a DNA methylation signature in blood cells from persons with down syndrome. Aging (Albany NY) **7**(2), 82 (2015)

Article IV

Computational Statistics
https://doi.org/10.1007/s00180-025-01630-5

ORIGINAL PAPER

A Newton-based variant of Exclusive Lasso for improved sparse solutions

Dayasri Ravi[1] · **Andreas Groll**[1]

Received: 12 December 2024 / Accepted: 7 April 2025

Abstract

Exclusive Lasso offers significant advantages in scenarios that require sparse solutions within groups, such as multi-omics or gene expression analysis. These applications involve inherent grouping structures where selecting only a subset of variables from each group is crucial due to high correlations among variables within groups. However, a key challenge in optimizing Exclusive Lasso stems from the non-differentiability of the L_1-norm within each group. To tackle this issue, we propose a method to transform this norm into a differentiable form using quadratic and sigmoid function approximations. This transformation facilitates the use of a straightforward Newton-based approach to solve the intricate optimization problem. Importantly, our proposed variant of Exclusive Lasso relaxes the strict requirement of selecting at least one variable per group, in contrast to the conventional Exclusive Lasso, and hence enables sparser solutions. Extensive simulation studies underscore the superior performance of our approach compared to both traditional Lasso methods and conventional Exclusive Lasso formulations.

Keywords Composite penalty · Newton-based approach · Quadratic approximation · Sigmoid function approximation · Variable selection

1 Introduction

Variable selection is a crucial step in the modeling process, particularly when dealing with high-dimensional datasets. Regularization techniques that induce structural sparsity have become popular tools for this purpose. One of the most prominent methods is the Lasso procedure (Tibshirani 1996), which employs the

✉ Dayasri Ravi
ravi@statistik.tu-dortmund.de

Andreas Groll
groll@statistik.tu-dortmund.de

[1] Department of Statistics, TU Dortmund University, Vogelpothsweg 87, 44221 Dortmund, NRW, Germany

Published online: 09 May 2025

L_1-norm to enforce sparsity, effectively selecting a subset of informative variables. Lasso has been widely adopted in various real-world applications, including microarray gene expression analysis (Güçkıran et al. 2019), cancer classification (Zheng and Liu 2011), sports prediction (Groll et al. 2015) and image classification (Huang et al. 2020).

Despite its popularity, Lasso has notable limitations. One major issue is its inability to consider the group structure of variables. Additionally, Lasso performs poorly when variables are highly correlated. Specifically, it struggles to select the correct features when informative variables are correlated with non-informative ones, leading to suboptimal performance (Zhao and Yu 2006). Several group regularizers have been developed to address the first limitation, among which the Group Lasso (Yuan and Lin 2006) is well-known. Group Lasso uses the $L_{2,1}$-norm regularization to induce sparsity between groups, i.e., applying the L_1-norm across groups to ensure sparsity at the group level and L_2-norm within each group such that coefficients within the group are regularized together.

However, Group Lasso also has its drawbacks, particularly in scenarios where groups of variables, such as those in multi-omics datasets, are highly correlated. In such cases, Group Lasso tends to select variables from only a few groups, often ignoring smaller or lower-dimensional groups like clinical variables, which are crucial for disease classification. This necessitates a method that can ensure variable selection from each group rather than selecting entire groups of variables.

The Exclusive Lasso was introduced to address these challenges, incorporating intra-group sparsity via an $L_{1,2}$-norm type penalty. This approach promotes sparsity within groups via the L_1-norm while relaxing sparsity between groups via L_2-norm, ensuring that at least one variable from each group is selected. Theoretical properties of this norm have been established in previous studies (Campbell and Allen 2017; Gregoratti et al. 2021), and its application is expanding to various fields, including multi-task feature learning (Zhou et al. 2010), image processing (Zhang et al. 2015), and clustering (Yamada et al. 2017).

Unlike Group Lasso, which can be solved using similar methods to Lasso by treating each group as an element (Liu et al. 2012), the Exclusive Lasso presents a more challenging optimization problem due to its composite penalty structure, combining the L_1-norm within groups. Various approaches have been proposed to tackle this problem, including coordinate descent and proximal gradient methods with soft-thresholding (Campbell and Allen 2017), dual Newton methods-based proximal point algorithms (Lin et al. 2020), and iterative re-weighted algorithms (Kong et al. 2014; Sun et al. 2020). An alternative approach converts the problem into a Lasso framework and solves it using a bisection algorithm, leveraging the piecewise linear property of Lasso (Sun et al. 2020). Recently, an optimization algorithm adopting the fast iterative shrinkage-thresholding algorithm (FISTA) has been proposed (Huang and Liu 2018).

We propose a novel approach to handle the $L_{1,2}$-norm-based penalty by using a smooth approximation, transforming the L_1-norm at the group level into a differentiable form. In linear regression models, this concept has already been utilized to simplify the optimization of least absolute values through a proposed estimator known as Smoothed Least-Absolute Deviations (Hitomi and Kagihara

2001). Our proposed method can accommodate any loss function, making it versatile for various applications. Once the norm is rendered continuous and differentiable, any Newton-based algorithm can be employed to solve the optimization problem efficiently. We utilize two specific approximations-the quadratic and sigmoid functions-to approximate the $L_{1,2}$-norm-based penalty. Newton-based optimization has been explored in the context of the L_1-norm through smoothing approximations in previous studies (Schmidt et al. 2007; Nkansah et al. 2021).

The main issue with the conventional Exclusive Lasso is that it enforces the selection of at least one variable from each group. Estimation is carried out using a coordinate descent algorithm, where variables are updated one at a time while others remain fixed. The consecutive variable is selected based solely on its correlation with the residuals, which must be equal to the correlation of other non-zero variables in its group. As a result, at least one non-zero variable is always chosen per group, even if none of the variables in that group are informative. In contrast, our approach addresses this by allowing for more flexible sparsity. Rather than strictly enforcing variable selection in each group, we let the penalty parameter control sparsity. By applying an additional, small threshold, we are able to nullify variable selection in the non-informative group. This avoids the need to select at least one variable per group, enabling the model to ignore uninformative groups more effectively. Hence, our proposed variant of Exclusive Lasso not only simplifies the optimization process but also retains the desirable properties of inducing sparsity within and between groups, ensuring robust variable selection across diverse datasets. Moreover, by applying the threshold in combination with using suitable starting values for the estimation scheme, we obtain more stable coefficient paths compared to conventional Exclusive Lasso. In this manuscript, we investigate the performance of our proposed methods through extensive simulation studies, focusing on scenarios that closely mimic real-world conditions where informative features are correlated and group allocation is random. We demonstrate that our model outperforms previous methods in similar simulation settings. Additionally, we apply the proposed method to two different real datasets to validate its practical utility.

The first real dataset involves gene expression data, where our goal is to identify the most significant genes within each module contributing to ovarian cancer. In this context, genes within a particular module or pathway are highly correlated, and each module may play a role in the disease's progression. The second dataset pertains to portfolio management, where stocks are diversified across different sectors to minimize risk. Here, the objective is to select the best stocks to achieve optimal returns while managing risk. In both cases, we highlight the importance of using Exclusive Lasso over Group Lasso.

The remainder of the manuscript is structured as follows. Section 2 introduces the Exclusive Lasso problem and our proposed estimation procedure. In Sect. 3, we present the simulation scenarios and compare our method with other Lasso procedures. The applicability of our model is demonstrated in Sect. 4 using the aforementioned application examples. Finally, Sect. 5 concludes.

2 Methodology

Let $\boldsymbol{x}_i \in \mathbb{R}^{p\times 1}$ be vector of covariates with the design matrix $\mathbf{X} = \left(\boldsymbol{x_1}, \ldots, \boldsymbol{x_n}\right)^\top \in \mathbb{R}^{n\times p}$ and $\mathbf{y} = \left(y_1, \ldots, y_n\right)^\top$ be the response vector. In principle, the intercept is not penalized and would normally be included as the first column in the design matrix. However, for simplicity, we assume the response is centered and exclude the intercept term. The estimation of the true parameter vector $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ in the linear model is typically achieved by maximizing the log-likelihood $l(\boldsymbol{\beta})$. Note that in regression models, this expression also depends on the design matrix $\mathbf{X}$ and the response vector $\mathbf{y}$, but these dependencies are omitted in this section for better readability. Following Campbell and Allen (2017), we consider that the indices of true parameter vector $\boldsymbol{\beta}$ are divided into non-overlapping groups. We consider the problem of the form

$$\min_{\boldsymbol{\beta}} f(\boldsymbol{\beta}) = -l(\boldsymbol{\beta}) + \frac{\lambda}{2} \sum_{g\in\mathcal{G}} \left(\sum_{k\in g} |\beta_k| \right)^2, \tag{1}$$

where $l(\boldsymbol{\beta})$ denotes the log-likelihood, $\lambda \geq 0$ is the penalty parameter and $\mathcal{G}$ is a collection of non-overlapping predefined groups $\{1, \ldots, G\}$ of indices g such that $\bigcup_{g\in\mathcal{G}} = \{1, \ldots, p\}$.

Let $S = \{1, \ldots, p\}$ be the set of indices corresponding to the entries of $\boldsymbol{\beta}$. We assume that each group contains at least one variable such that $S \cap g \neq \emptyset$, for all $g \in \mathcal{G}$. Let $d = |S \cap g|$ and we construct square matrices $\mathbf{M}_g$ with dimensions $d \times d$ such that

$$\mathbf{M}_{g_{ij}} = \begin{cases} 1, \text{ if } i = 1, \\ 0, \text{ if } i > 1. \end{cases}$$

Let $\mathbf{b} \in \mathbb{R}^{p\times 1}$ represent the vector of absolute coefficient values, i.e. $\mathbf{b} := \left(|\beta_1|, \ldots, |\beta_p|\right)^\top$. Now the penalty term from Eq. (1) can be rewritten as

$$P(\boldsymbol{\beta}) = \frac{1}{2}(\mathbf{Mb})^\top \mathbf{Mb}\,, \tag{2}$$

where $\mathbf{M} \in \mathbb{R}^{p\times p}$ is a block diagonal matrix built from matrices $\mathbf{M}_1, \ldots, \mathbf{M}_G$.

We use smooth approximations for the transformation of the L_1-norm into a continuous and differentiable function. Below, we propose two approximations through quadratic and sigmoid functions, allowing us to derive the gradient and Hessian of the penalty term $P(\boldsymbol{\beta})$. We then provide an additional thresholding technique to select the optimal variables from each group in Sect. 2.3.

2.1 Quadratic approximation

Following Oelker and Tutz (2017), we use a quadratic approximation for the transformation of the L_1-norm into a differentiable function given by

$|x| \approx \sqrt{x^2 + c}$. Here, c is a small positive number and controls how closely the approximation matches the L_1-norm. It acts as a second layer tuning parameter, which typically is not tuned but just fixed to a rather small value. In practice, a value of $c \approx 10^{-5}$ performs optimally (Oelker and Tutz 2017). With this approximation, the penalty term from Eq. (1) can be rewritten as

$$P(\boldsymbol{\beta}) = \frac{1}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} \sqrt{\beta_k^2 + c} \right)^2 .$$

The vector $\mathbf{b}$ can be rewritten as $\mathbf{b} \approx \left(\sqrt{(\beta_1^2 + c)}, \ldots, \sqrt{(\beta_p^2 + c)} \right)^\top$, which we will used in this section. Then, one obtains

$$\frac{\partial \mathbf{b}}{\partial \beta_k} = \frac{\beta_k}{\sqrt{(\beta_k^2 + c)}} \quad \& \quad \frac{\partial^2 \mathbf{b}}{\partial \beta_k^2} = \frac{c}{\left(\beta_k^2 + c\right)^{\frac{3}{2}}} . \tag{3}$$

2.2 Sigmoid function approximation

We rewrite the absolute value function as

$$|x| = \max(x, 0) + \max(-x, 0).$$

We approximate the plus function $(x)_+ = \max(x, 0)$ by the integral of a sigmoid function (Chen and Mangasarian 1996):

$$(x)_+ \approx p(x, c) = x + \frac{1}{c} log\left(1 + e^{(-cx)}\right).$$

Similarly, we have

$$(-x)_+ \approx p(-x, c) = -x + \frac{1}{c} log\left(1 + e^{(cx)}\right).$$

Now combining $p(x, c)$ and $p(-x, c)$, we can write the smoothing approximation of $|x|$ as the sum of the integral of two sigmoid functions given by

$$\begin{aligned} |x| &\approx p(x, c) + p(-x, c) \\ &= \frac{1}{c} \left[\left(log\left(1 + e^{(-cx)}\right) + log\left(1 + e^{(cx)}\right) \right] \\ &=: |x|_c . \end{aligned}$$

One can show that $\lim_{x \to \infty} |x|_c = |x|$. The proof is similar to that in Lee and Mangasarian (2001), and we direct readers to it for further details. We rewrite vector $\mathbf{b}$ as $\mathbf{b} \approx \left[\left(log\left(1 + e^{(-c\beta_k)}\right) + log\left(1 + e^{(c\beta_k)}\right) \right] \right]_{1 \le k \le p}^\top$, which we will used in this section.

The penalty term from Eq. (1) can be rewritten as

$$P(\boldsymbol{\beta}) = \frac{1}{2}\sum_{g\in\mathcal{G}}\left(\sum_{k\in g}\left[\left(log\left(1+e^{(-c\beta_k)}\right)+log\left(1+e^{(c\beta_k)}\right)\right]\right)^2.$$

Hence, we have

$$\begin{aligned}\frac{\partial \mathbf{b}}{\partial \beta_k} =& \frac{1}{c}\left(\frac{-ce^{-c\beta_k}}{1+e^{-c\beta_k}}+\frac{ce^{c\beta_k}}{1+e^{c\beta_k}}\right)\\ =& \frac{\left(1+e^{c\beta_k}\right)-\left(1+e^{-c\beta_k}\right)}{\left(1+e^{-c\beta_k}\right)\left(1+e^{c\beta_k}\right)}\\ =& \left(1+e^{-c\beta_k}\right)^{-1}-\left(1+e^{c\beta_k}\right)^{-1}.\end{aligned} \tag{4}$$

$$\begin{aligned}\frac{\partial^2 \mathbf{b}}{\partial \beta_k^2} =& ce^{-c\beta_k}\left(1+e^{-c\beta_k}\right)^{-2}+ce^{c\beta_k}\left(1+e^{c\beta_k}\right)^{-2}\\ =& ce^{c\beta_k}\left(1+e^{c\beta_k}\right)^{-2}\left[\left(e^{-c\beta_k}\right)^2\frac{\left(1+e^{c\beta_k}\right)^2}{\left(1+e^{-c\beta_k}\right)^2}+1\right]\\ =& 2ce^{c\beta_k}\left(1+e^{c\beta_k}\right)^{-2}.\end{aligned} \tag{5}$$

In contrast to the smaller value of the approximation parameter c chosen in Sect. 2.1, here we need to select a larger value of c for an optimal solution. Hence, along the lines of Schmidt et al. (2007), we introduce a more stable approach using a continuous strategy. Instead of empirically choosing the approximation parameter c, we take Newton's steps, starting from a small c value where the approximation is appropriate and stopping at a sufficiently large value of c. This approach is expected to require less iterations than other unconstrained approximation methods.

2.3 Newton method-$L_{1,2}$ sparsity algorithm

Using one of the approximations of vector $\mathbf{b}$ described in either Sect. 2.1 or 2.2, we apply matrix and vector differentiation rules to derive the gradient and Hessian of the penalty term $P(\boldsymbol{\beta})$ from Eq. (2) as

$$\begin{aligned}\frac{\partial}{\partial \beta_k}P(\boldsymbol{\beta}) &= \frac{\partial \mathbf{b}}{\partial \beta_k}\cdot\left(\mathbf{M}^\top\mathbf{M}\mathbf{b}\right)_k,\\ \frac{\partial^2}{\partial \beta_{kl}}P(\boldsymbol{\beta}) &= \begin{cases}\left[\frac{\partial^2 \mathbf{b}}{\partial \beta_k^2}\cdot\left(\mathbf{M}^\top\mathbf{M}\mathbf{b}\right)_k+\frac{\partial \mathbf{b}}{\partial \beta_k}\cdot\sum_{l=1}^{p}(\mathbf{M}^\top\mathbf{M})_{kl}\cdot\frac{\partial \mathbf{b}}{\partial \beta_l}\right], & k=l,\\ \frac{\partial \mathbf{b}}{\partial \beta_k}\cdot(\mathbf{M}^\top\mathbf{M})_{kl}\cdot\frac{\partial \mathbf{b}}{\partial \beta_l}, & k\neq l.\end{cases}\end{aligned} \tag{6}$$

With Eq. (6), we obtain the gradient vector $\nabla P(\boldsymbol{\beta})$ and the Hessian matrix $\nabla^2 P(\boldsymbol{\beta})$ of the penalty term where

$$\nabla P(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial P(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial P(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial P(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix}. \tag{7}$$

$$\nabla^2 P(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_1^2} & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2} & \cdots & \frac{\partial^2 P(\boldsymbol{\beta})}{\partial \beta_p^2} \end{pmatrix}. \tag{8}$$

Hence, the full gradient $\mathbf{g}$ and Hessian $\mathbf{H}$ of our Exclusive Lasso problem from Eq. (1) yield $\mathbf{g} = -\nabla l(\boldsymbol{\beta}) + \lambda \cdot \nabla P(\boldsymbol{\beta})$ and $\mathbf{H} = \nabla^2 l(\boldsymbol{\beta}) + \lambda \cdot \nabla^2 P(\boldsymbol{\beta})$, respectively. The penalty term is hence differentiable and we perform Newton iterations of the form $\boldsymbol{\beta}^{(t+1)} := \boldsymbol{\beta}^{(t)} - s \cdot (\nabla^2 g(\boldsymbol{\beta}^{(t)}))^{-1} \nabla g(\boldsymbol{\beta}^{(t)})$, where $g(\boldsymbol{\beta})$ is a continuous and twice-differentiable approximation function of $f(\boldsymbol{\beta})$ and s denotes the step size.

Applying an L_2-norm between groups prevents entire groups of coefficients from being set to zero. However, this approach is problematic when a group is non-informative, as L_2 regularization only shrinks the coefficients without completely eliminating them. As a result, non-informative groups may still influence the model, potentially reducing model performance. To address this scenario more effectively, we use a small threshold to automatically set covariates estimated below it to zero, thereby nullifying the representation of non-informative groups. This is not a tuning parameter but a small approximation value (10^{-4} has worked well in our experience for most problems), ensuring that only estimates very close to zero are set to zero to nullify the group's effect. Since we standardize the variables to a variance of one before fitting the model, a fixed (and rather small) threshold remains reasonable and provides consistent behavior across different datasets. Additionally, this parameter can be turned off if representation from each group is preferred. This also results in more stable coefficient paths compared to the conventional Exclusive Lasso.

2.4 Choice of approximation

We present two algorithms based on the chosen approximation, either a quadratic or sigmoid function. First, we perform a simple Newton method with quadratic approximation from Sect. 2.1 and a chosen step size of $s = 1$ in Algorithm 1. The advantage of this method is that we do not have to tune the approximation parameter c. However, this algorithm tends to be slower due to the number of required iterations (see Sect. 3.3). The NM-$L_{1,2}$ sparsity algorithm with sigmoid function approximation is shown in Algorithm 3 with gradient and Hessian calculated using Algorithm 2. Here, however, we choose to tune the approximation parameter c as no clear empirical value has been stated by the previous studies. Additionally, we

choose step size s using a back-tracking line search satisfying the Armijo condition (Armijo 1966), i.e.

$$g(\boldsymbol{\beta}^{(t)} + s^{(t)}d^{(t)}) \leq g(\boldsymbol{\beta}^{(t)}) + q \cdot s^{(t)}\nabla g(\boldsymbol{\beta}^{(t)})^{\top}d^{(t)},$$

where $g(\boldsymbol{\beta}^{(t)})$ is the objective function, $d^{(t)}$ is the search direction at the current iteration t, and q is the decrease parameter. We employ cubic interpolation to estimate the objective function along the search direction by fitting a cubic polynomial to the function values at the current iteration. The minimum of the interpolated cubic polynomial is subsequently determined to identify the step size estimate that minimizes the objective function along the search direction. We choose a sufficiently smaller decrease parameter $q = 10^{-4}$ to ensure that the step size chosen in the line search provides a sufficient decrease in the objective function.

Algorithm 1 NM-$L_{1,2}$ Sparsity Algorithm with Quadratic Approximation

Input: Design matrix $\mathbf{X} \in \mathbb{R}^{n\times p}$, response vector $\mathbf{y} \in \mathbb{R}^{n\times 1}$, regularization parameter λ, starting values $\boldsymbol{\beta}_0$, group matrix $\mathbf{M} \in \mathbb{R}^{p\times p}$, tolerance ϵ, maximum iterations N
Output: Estimated parameter vector $\hat{\boldsymbol{\beta}}$
Initialize $\boldsymbol{\beta} = \boldsymbol{\beta}_0$
Set $c = 10^{-5}$, $i = 0$
Compute $\mathbf{C} = \mathbf{M}^{\top}\mathbf{M}$, $\mathbf{D} = \mathbf{C} + \mathbf{C}^{\top}$
while $i < N$ **do**
    Compute $\frac{\partial \mathbf{b}}{\partial \beta_k}$ & $\frac{\partial^2 \mathbf{b}}{\partial \beta_k^2}$, $\forall\ k \in g, g \in \mathcal{G}$ from Eq. (3)
    Compute gradient and Hessian for the penalty term $\nabla P(\boldsymbol{\beta})$ & $\nabla^2 P(\boldsymbol{\beta})$ from Eq. (7) and Eq. (8)
    Compute gradient: $\mathbf{g} = -\nabla l(\boldsymbol{\beta}) + \lambda \cdot \nabla P(\boldsymbol{\beta})$
    Compute Hessian: $\mathbf{H} = \nabla^2 l(\boldsymbol{\beta}) + \lambda \cdot \nabla^2 P(\boldsymbol{\beta})$
    Update parameter vector: $\boldsymbol{\beta}_{new} = \boldsymbol{\beta} - \mathbf{H}^{-1}\mathbf{g}$
    Check convergence: if $\|\boldsymbol{\beta}_{new} - \boldsymbol{\beta}\|_2 < \epsilon$; then stop
    Increase i to $i + 1$
    $\boldsymbol{\beta} = \boldsymbol{\beta}_{new}$
end while
Return $\boldsymbol{\beta}$

2.5 Software implementation

The NM-$L_{1,2}$ sparsity algorithm is coded and implemented in `Python` (version 3.12.3). The source code is available on GitHub (URL: https://github.com/draviis/NML12.git).

Algorithm 2 GradHessian

Input: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, response vector $\mathbf{y} \in \mathbb{R}^{n \times 1}$, parameter vector $\boldsymbol{\beta}$, regularization parameter λ, group matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, approximation parameter c
Output: Negative log-likelihood $-l(\boldsymbol{\beta})$, gradient $\mathbf{g}$, Hessian $\mathbf{H}$
Compute $\mathbf{C} = \mathbf{M}^\top \mathbf{M}$, $\mathbf{D} = \mathbf{C} + \mathbf{C}^\top$
Compute $\frac{\partial \mathbf{b}}{\partial \beta_k}$ & $\frac{\partial^2 \mathbf{b}}{\partial \beta_k^2}$, $\forall\ k \in g, g \in \mathcal{G}$ from Eq. (4) and Eq. (5)
Compute gradient and Hessian for the penalty term $\nabla P(\boldsymbol{\beta})$ & $\nabla^2 P(\boldsymbol{\beta})$ from Eq. (7) and Eq. (8)
Compute gradient: $\mathbf{g} = -\nabla l(\boldsymbol{\beta}) + \lambda \cdot \nabla P(\boldsymbol{\beta})$
Compute Hessian: $\mathbf{H} = \nabla^2 l(\boldsymbol{\beta}) + \lambda \cdot \nabla^2 P(\boldsymbol{\beta})$
Return $-l(\boldsymbol{\beta})$,$\mathbf{g}$,$\mathbf{H}$

Algorithm 3 NM-$L_{1,2}$ Sparsity Algorithm with Sigmoid Function Approximation

Input: Design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, response vector $\mathbf{y} \in \mathbb{R}^{n \times 1}$, parameter vector $\boldsymbol{\beta}$, regularization parameter λ, starting values $\boldsymbol{\beta}_0$, group matrix $\mathbf{M} \in \mathbb{R}^{p \times p}$, tolerance ϵ, maximum iterations N, initial c_0, $c_{\max}$, update parameters η_1, η_2
Output: Estimated parameter vector $\hat{\boldsymbol{\beta}}$
Initialize $\boldsymbol{\beta} = \boldsymbol{\beta}_0$, $c = c_0$, $s = 1$
Evaluate $f, \mathbf{g}, \mathbf{H} \leftarrow \text{GradHessian}(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}, \lambda, \mathbf{M}, c)$
Set $i = 0$
while $i < N$ **do**
    $\mathbf{d} = -\mathbf{H}^{-1}\mathbf{g}$
    $\boldsymbol{\beta}_{\text{new}} = \boldsymbol{\beta} + s\mathbf{d}$
    Evaluate $f_{\text{new}}, \mathbf{g}_{\text{new}} \leftarrow \text{GradHessian}(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}_{\text{new}}, \lambda, \mathbf{M}, c)$
    Armijo Backtracking Line Search
    while $f_{\text{new}} > f + 10^{-4} \cdot s \cdot \mathbf{g}^\top \mathbf{d}$ **do**
        Adjust step size s using cubic interpolation
        $\boldsymbol{\beta}_{\text{new}} \leftarrow \boldsymbol{\beta} + s\mathbf{d}$
        Evaluate $f_{\text{new}}, \mathbf{g}_{\text{new}} \leftarrow \text{GradHessian}(\mathbf{X}, \mathbf{y}, \boldsymbol{\beta}_{\text{new}}, \lambda, \mathbf{M}, c)$
    end while
    Continuation Strategy
    if Armijo line search required multiple steps **then**
        $c \leftarrow \min(c \cdot \eta_1, c_{\max})$
    else
        $c \leftarrow \min(c \cdot \eta_2, c_{\max})$
    end if
    Check convergence: if $\|\boldsymbol{\beta}_{new} - \boldsymbol{\beta}\|_2 < \epsilon$ & $c = c_{\max}$; then stop
    Increase i to $i + 1$
    $\boldsymbol{\beta} = \boldsymbol{\beta}_{\text{new}}$; $f = f_{\text{new}}$; $\mathbf{g} = \mathbf{g}_{\text{new}}$
end while
Return $\boldsymbol{\beta}$

3 Simulations

In this section, we provide a comprehensive simulation study to evaluate the performance of our method across various scenarios.

3.1 Setting

We simulate examples using the model $\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$, where $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, 1)$ with $n = 100$ observations and $p = 100$ variables. Let p_s be the number of informative, i.e., signal variables out of the p variables. Following Sun et al. (2020) and Campbell and Allen (2017), we consider multivariate Gaussian covariates with the following types of covariance structures:

- Example 1: We consider the informative variables to be pairwise correlated with $\rho = 0.8$.
- Example 2: The data is simulated from a multivariate normal distribution with a Toeplitz covariance matrix $\boldsymbol{\Sigma}$ with entries $\Sigma_{i,j} = a^{|i-j|}$ for variables in the same group, and $\Sigma_{i,j} = b^{|i-j|}$ for variables in different groups. We use a moderate correlation of $a = 0.6$ and $b = 0.3$, resulting in a high correlation within and a low correlation between groups.

As suggested by Sun et al. (2020), the Exclusive Lasso performs well if the group contains at least one informative variable. However, when many informative variables are placed in the same group, the probability of selecting any single informative variable decreases. We allocate each informative variable along with its correlated variables in a group and consider this ideal situation. We set the number of groups as the number of informative variables. This optimal arrangement is termed "fixed allocation." However, we also consider the scenario closer to real-world scenarios with random allocation of groups. We refer to this as "random allocation." We choose the number of informative variables to be $p_s = 10, 20, 40$ with the number of groups as 5, 10, and 20, respectively. For random allocation, we consider the number of groups to be 25. We choose the regularization parameter λ through 5-fold cross-validation in all the above scenarios.

We report the results in terms of the variable selection accuracy and

$$\text{F1 score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}},$$

where

$$\text{Precision} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$
$$\text{Recall} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}.$$

The F1 score (Van Rijsbergen 1979) represents the harmonic mean of precision and recall, which accounts for both false positives and false negatives. The metric ranges from 0 to 1, where a higher value indicates a better balance between precision and recall, hence an overall effective classification model.

3.2 Results

We compare *NM-*$L_{1,2}$ with Lasso (Friedman et al. 2010), conventional Exclusive Lasso with coordinate descent from the `Exclusive Lasso` R package (URL: https://github.com/DataSlingers/ExclusiveLasso.git), as well as Group Lasso (Simon et al. 2019).

We examine the F1 scores over 50 random training datasets across two distinct examples explained above. The results are visually represented through boxplots in Figs. 1 and 2, which illustrate the performance across different numbers of informative features (p_s) under both fixed and random allocation scenarios.

The comparison shows that both our NM-$L_{1,2}$ algorithms perform much better than Lasso and Group Lasso. Notably, Lasso and Group Lasso demonstrate sub-optimal performance, frequently selecting only a limited number of variables. This under-performance can be due to the inherent correlations within groups and the

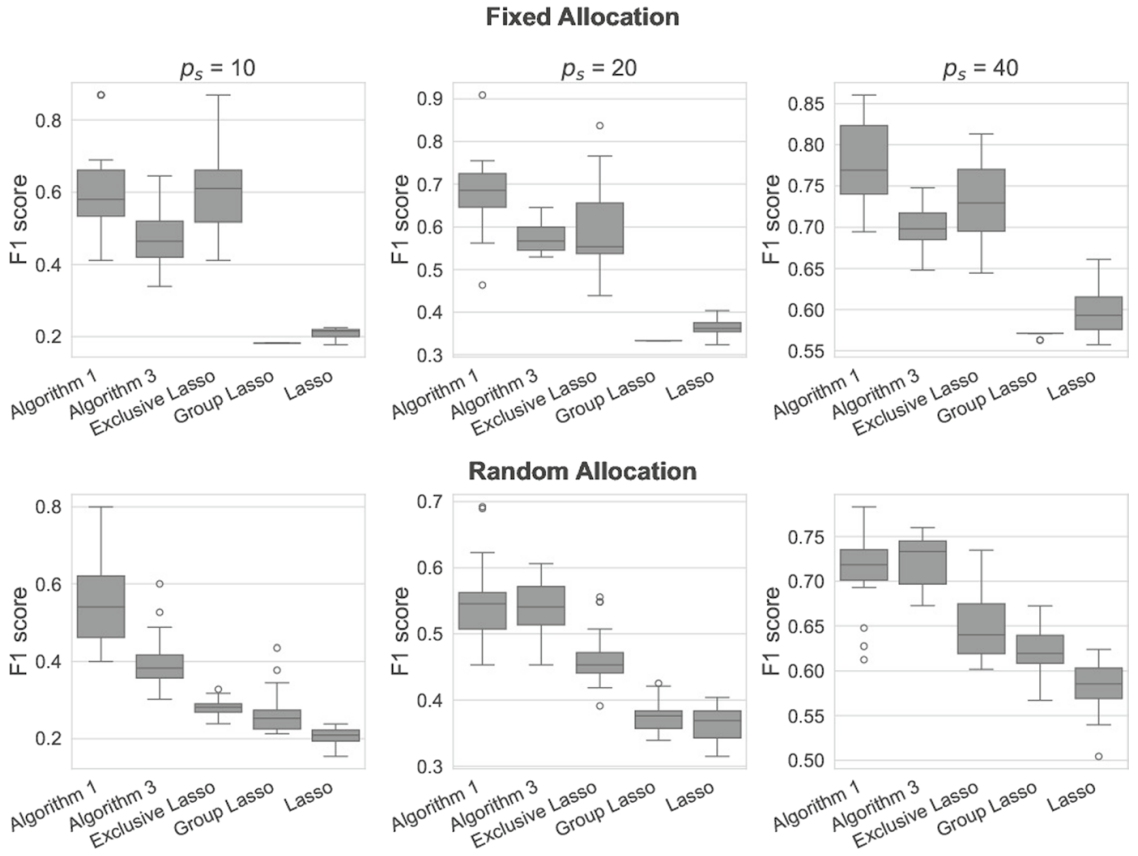

Fig. 1 Boxplots of F1 scores from Example 1 across different numbers of informative features (p_s). *Top row*: Fixed allocation *Bottom row*: Random allocation

Table 1 Variable selection accuracy (standard errors in brackets) with fixed allocation averaged over 50 training datasets; best-performing modeling approach per setting in bold font

Example	p_s	Algorithm 1	Algorithm 3	Exclusive Lasso	Group Lasso	Lasso
Example 1	10	**0.87** (0.013)	0.78 (0.014)	0.87 (0.015)	0.10 (0.000)	0.29 (0.005)
	20	**0.82** (0.021)	0.72 (0.011)	0.75 (0.027)	0.20 (0.000)	0.36 (0.010)
	40	**0.79** (0.018)	0.67 (0.011)	0.73 (0.024)	0.40 (0.001)	0.51 (0.012)
Example 2	10	**0.89** (0.009)	0.80 (0.012)	0.85 (0.012)	0.10 (0.000)	0.33 (0.007)
	20	**0.81** (0.014)	0.73 (0.014)	0.78 (0.014)	0.20 (0.000)	0.41 (0.006)
	40	**0.75** (0.017)	0.65 (0.007)	0.73 (0.015)	0.40 (0.001)	0.55 (0.010)

Table 2 Variable selection accuracy (standard errors in brackets) with random allocation averaged over 50 training datasets; best-performing modeling approach per setting in bold font

Example	p_s	Algorithm 1	Algorithm 3	Exclusive Lasso	Group Lasso	Lasso
Example 1	10	**0.85** (0.013)	0.70 (0.014)	0.53 (0.008)	0.42 (0.027)	0.29 (0.006)
	20	**0.70** (0.012)	0.68 (0.009)	0.58 (0.012)	0.34 (0.013)	0.35 (0.007)
	40	**0.72** (0.012)	0.70 (0.009)	0.62 (0.014)	0.52 (0.011)	0.49 (0.008)
Example 2	10	**0.73** (0.019)	0.63 (0.016)	0.54 (0.017)	0.43 (0.036)	0.32 (0.009)
	20	**0.68** (0.017)	0.64 (0.011)	0.56 (0.014)	0.36 (0.020)	0.41 (0.010)
	40	**0.68** (0.015)	0.63 (0.010)	0.62 (0.013)	0.47 (0.009)	0.54 (0.008)

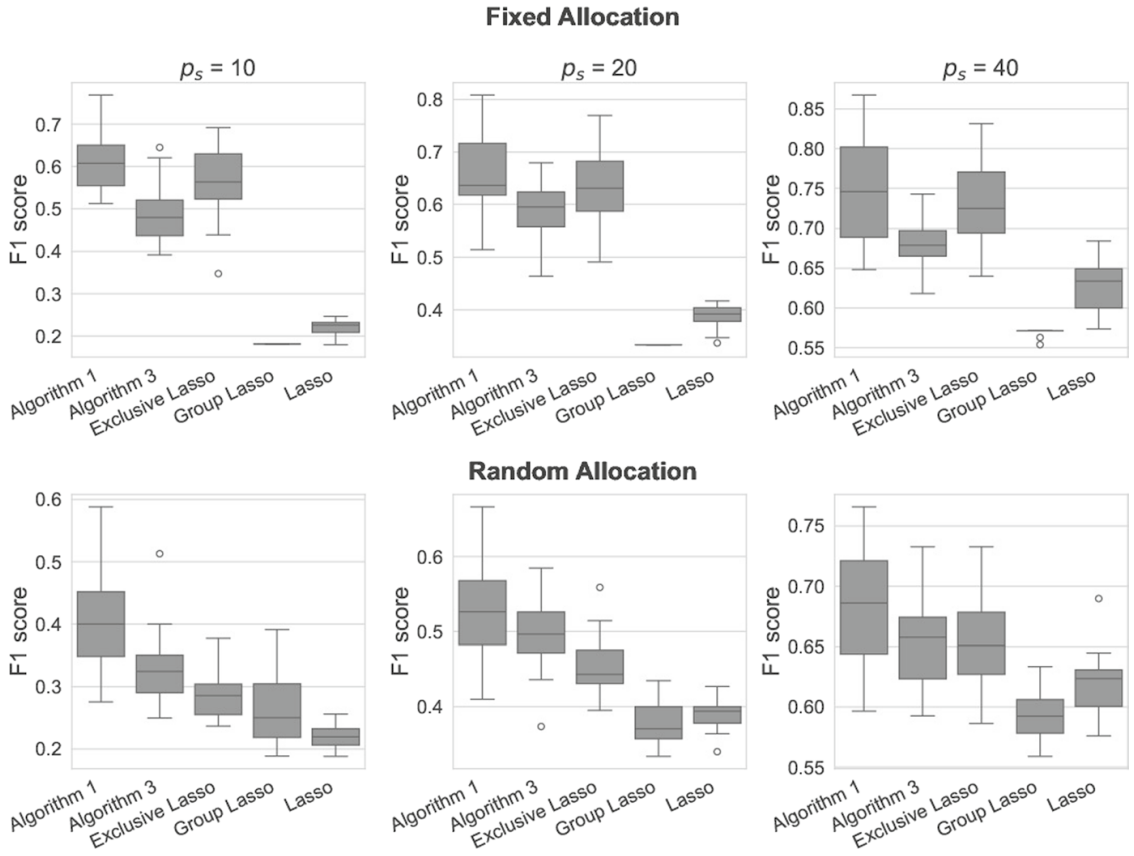

Fig. 2 Boxplots of F1 scores from Example 2 across different numbers of informative features (p_s). *Top row*: Fixed allocation *Bottom row*: Random allocation

inter-group sparsity enforced by Group Lasso, which leads to the selection of irrelevant features from the same group.

Under fixed allocation conditions, Algorithm 1 and Exclusive Lasso demonstrate comparable performance, achieving similar F1 scores. However, Algorithm 1 demonstrates a significant advantage over Exclusive Lasso, particularly in situations involving a high number of correlated features within a group (e.g., $p_s = 20, 40$). Although Example 2 does not reflect the typical scenarios where Exclusive Lasso would be employed, our model still demonstrates better overall performance than other methods. This enhanced performance of NM-$L_{1,2}$ is attributed to its superior tolerance to within-group correlations. Additionally, the comparison shows substantial performance differences between fixed and random allocation. Algorithm 3 shows sub-optimal performance under fixed allocation conditions but outperforms the traditional Exclusive Lasso in random allocation scenarios. The Exclusive Lasso is structured to select at least one variable from each group, regardless of its informativeness. This rigid selection approach leads to poor performance when a group either contains multiple informative variables that are correlated or lacks any informative variables altogether, often due to random group allocation.

The variable selection accuracy, along with the standard errors (SEs), averaged over the same 50 training datasets, are presented in Tables 1, 2. Consistent with the F1 results, Lasso and Group Lasso exhibit poor performance. Although Exclusive Lasso shows an average performance in certain scenarios, NM-$L_{1,2}$ outperforms it, especially under random allocation conditions.

Algorithm 1 consistently delivers the best performance in variable selection. In contrast, Algorithm 3 outperforms its competitors only in random allocation scenarios, demonstrating its robustness in handling datasets with correlated features, which are common in real-world applications.

3.3 Runtime analysis

We compare the runtime performance of the NM-$L_{1,2}$ algorithm using two different approximation methods: quadratic approximation (Algorithm 1) and sigmoid function approximation (Algorithm 3). We generate data based on Example 1, considering two scenarios with $p = 100$ and $p = 500$, while keeping the number of observations fixed at $n = 100$. The variables are divided into 5 groups with a fixed allocation.

Figure 3 presents the average runtimes of the two algorithms over 50 random iterations. Our findings indicate a significant difference in computational performance between the two methods. Algorithm 3, which employs the sigmoid function approximation, demonstrates consistently faster runtimes compared to Algorithm 1, regardless of the increase in dimensionality. This efficiency can be attributed to the computational simplicity of the sigmoid function compared to the more complex quadratic function.

Algorithm 1, which relies on quadratic approximation, exhibits a longer convergence time. The increased computational burden associated with solving

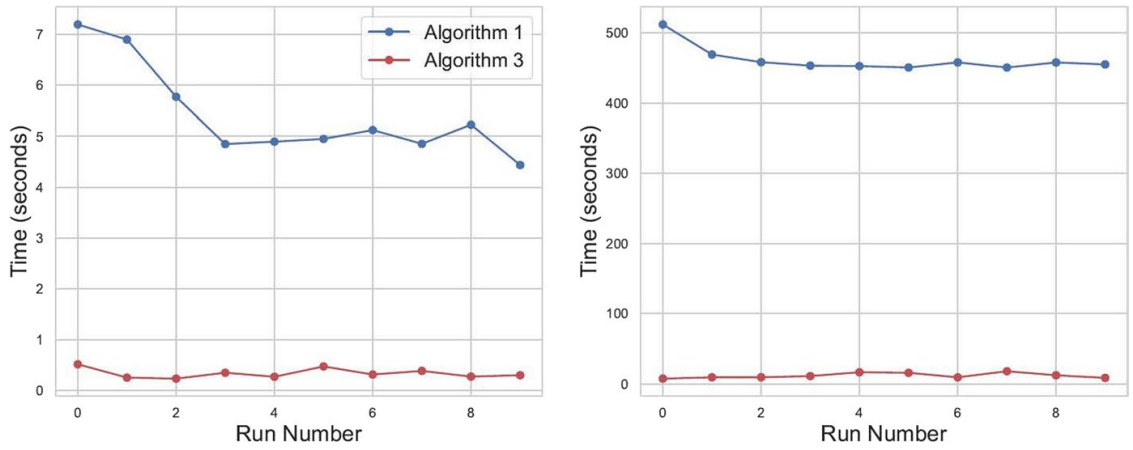

Fig. 3 Runtime comparison for different values of p

the quadratic function results in substantially higher runtimes, particularly as the problem dimensionality increases.

It is important to note that we did not include the runtime analysis for the conventional Exclusive Lasso implemented with coordinate descent from the `Exclusive Lasso` R package (URL: https://github.com/DataSlingers/ExclusiveLasso.git). The package is optimized using `Rcpp` (Eddelbuettel and Balamuta 2018), a tool that integrates C++ code within R. Given that `Rcpp` typically provides a performance advantage over pure Python implementations, including it in our runtime comparison could introduce an unfair bias favoring the R implementation. The exclusion of the conventional Exclusive Lasso from our runtime comparison ensures a fair assessment by avoiding potential biases introduced by differing underlying programming languages and optimizations.

4 Applications

In the following section, we apply our method to real-world datasets from finance and biology. The regularization parameter λ is tuned via cross-validation. As Algorithm 3 performs better in random allocations and is faster, as demonstrated in the previous section, we refer to Algorithm 3 as NM-$L_{1,2}$ in the following section.

4.1 Gene expression analysis

We study the two ovarian cancer datasets (OC) that were retrieved from the Gene Expression Omnibus (GEO) database (URL: https://www.ncbi.nlm.nih.gov/geo/) using the "GEOquery" R/`Bioconductor` library with the GEO accession GSE51373 (Koti et al. 2013) and GSE28739 (Trinh et al. 2011). The first dataset contains the gene expression profiles of 54 675 genes from a total of 28 samples, including 12 samples from the resistant cohort and 16 samples from the sensitive cohort. The second dataset contains the gene expression profiles of 16 096 genes from a total of 50 samples, including 30 samples from the resistant cohort and 20

samples from the sensitive cohort. Samples from the resistant cohort are designated as class 0, while sensitive samples are designated as class 1. This dichotomized classification is used as the response variable in our model.

To identify groups of genes for our analysis, we utilize the Weighted Gene Co-Expression Network Analysis (WGCNA) method (Zhang and Horvath 2005), a network-based systems biology approach. WGCNA clusters genes into modules according to their expression patterns, with genes within a module showing high levels of co-expression. This high degree of co-expression suggests that these genes are likely involved in similar biological functions or pathways, thereby allowing us to identify distinct groups of genes with correlated expression profiles.

We follow the pre-processing steps outlined in Wang et al. (2018), selecting only the top one percentile of genes. After standardization, WGCNA is performed on the genes from the filtered-resistant cohort, identifying 16 and 18 gene modules for OC dataset 1 and OC dataset 2, respectively. Detailed analysis is presented in the supplementary material of this manuscript.

We compare the variable selection performance of our method with a sigmoid approximation against the three other methods previously discussed. For all four models, we use tuned regularization parameters selected from 10-fold cross-validation. The selected variables from each model are input into a support vector machine (SVM) with a linear kernel (Kecman 2005). To avoid introducing split bias, we employ 10-fold cross-validation instead of a train-test split. The average classification accuracy and F1 scores of the models are presented in Tables 3, 4.

Figure 4 displays the percentages of genes selected from each module as clustered by the WGCNA model (see supplementary material). For OC dataset 1, Group Lasso selects genes predominantly from a single module due to inter-group sparsity. However, this approach leads to poor variable selection and results in low classification accuracy (see Tables 3 and 4). Since genes are highly correlated within each module, it is beneficial to employ intra-group sparsity rather than inter-group sparsity to avoid selecting all highly correlated, redundant genes. It is essential to select

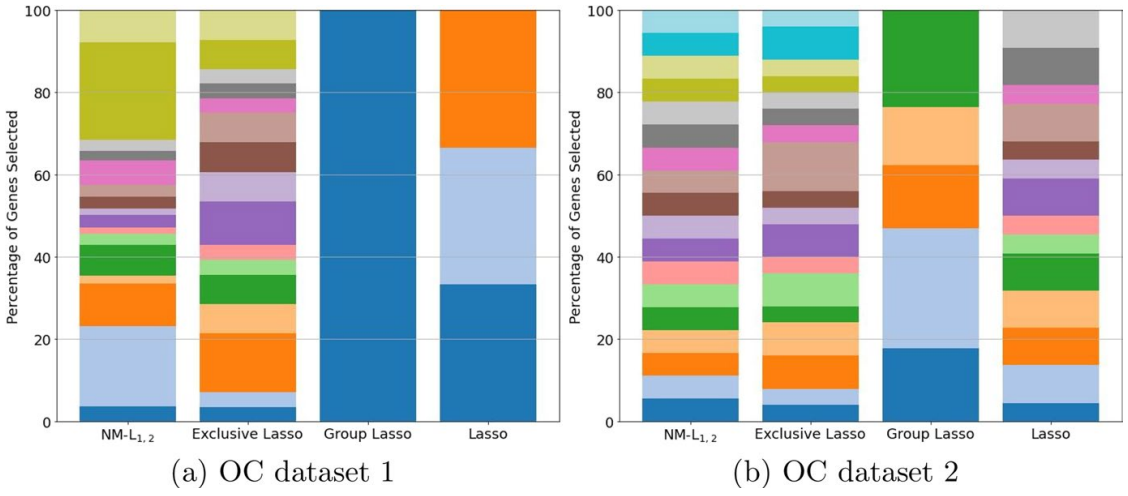

Fig. 4 Comparison of gene selection percentages for the two OC datasets

Table 3 OC dataset 1: Mean accuracy and mean F1 score; best-performing modeling approach per setting in bold font

Method	Top 3 genes		Top 5 genes	
	Mean accuracy	Mean F1 score	Mean accuracy	Mean F1 score
NM-$L_{1,2}$	**0.87**	**0.84**	**0.93**	**0.92**
Exclusive Lasso	0.83	0.82	0.80	0.76
Group Lasso	0.68	0.63	0.72	0.66
Lasso	0.87	0.84	0.87	0.85

Table 4 OC dataset 2: Mean accuracy and mean F1 score; best-performing modeling approach per setting in bold font

Method	Top 3 genes		Top 5 genes	
	Mean accuracy	Mean F1 score	Mean accuracy	Mean F1 score
NM-$L_{1,2}$	**0.92**	**0.89**	**0.94**	**0.93**
Exclusive Lasso	0.84	0.79	0.84	0.79
Group Lasso	0.62	0.52	0.70	0.63
Lasso	0.76	0.72	0.86	0.82

genes from all modules, as genes in different modules often participate in distinct biological functions or pathways.

Although Lasso performs better in terms of accuracy on OC dataset 1, it selects genes from only three modules, thereby ignoring potentially significant variables from other modules. In contrast, both our model and Exclusive Lasso select variables from every module, though the proportions of genes selected from each module vary. Notably, our model achieves the best accuracy and F1 score with as few as three genes in both datasets.

In this application study, we did not analyze the selected genes or assess their biological significance. Our primary objective was to demonstrate the use case and compare the performance of different Lasso models. However, the importance of the selected genes has been documented in previous cancer studies. For example, among the variables selected by our model, DDX39B has been shown to sensitize ovarian cancer cells (Xu et al. 2020), and the antioxidant enzyme CAT is associated with cancer survival risks (Belotte et al. 2015). Similarly, the role of Tensin in cancer research is well established (Mainsiouw et al. 2023).

4.2 Exchange-traded fund (ETF)

The exchange-traded fund (ETF) data has been considered in the context of intra-group sparsity in a previous study (Lin et al. 2020). We download the daily stock price data in the US market from '2023-01-01' to '2023-12-31' from Yahoo Finance (URL: https://finance.yahoo.com/). There are 2057 stocks grouped into 12 sectors in

Table 5 Number of stocks in each US sector

Sector	Number of stocks
Basic materials	158
Communication services	61
Consumer cyclical	208
Consumer defensive	91
Energy	176
Financial services	589
Healthcare	135
Industrials	273
Real Estate	162
Technology	131
Utilities	73

the US market to diversify risks (see Table 5). We choose the historical daily return of the S&P 500 index as our response variable. We aim to select the best set of stocks from each sector to predict the S&P 500 index most effectively.

We use 10-fold cross-validation and report the average Mean Squared Error (MSE) in Table 6. We could not run the Group Lasso model due to computational problems, as the coordinate descent did not converge even with the maximum number of iterations. Therefore, we exclude this model from our analysis. We do not observe much difference in the losses among the three methods. However, as shown in Fig. 5, Lasso requires a larger number of variables to reach the minimum loss, whereas our method selects the minimum set of variables to achieve the best returns for the S&P 500 index.

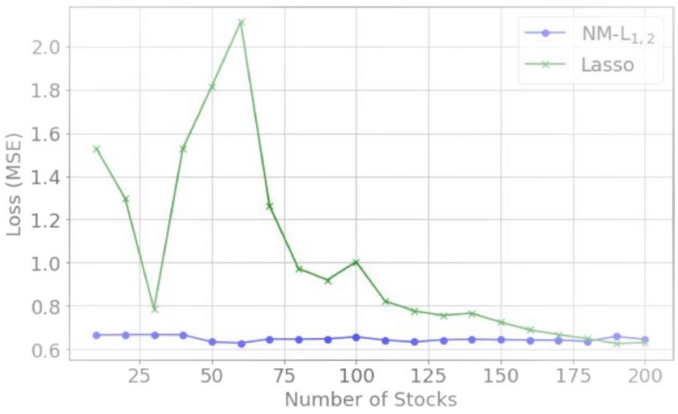

Fig. 5 Comparison of MSE with respect to the number of variables for NM-$L_{1,2}$ and Lasso

Table 6 Average MSE (standard errors in brackets) of models; best-performing modeling approach per setting in bold font

Model	MSE (SE)
NM-$L_{1,2}$	**0.61** (0.059)
Exclusive Lasso	0.63 (0.053)
Lasso	0.68 (0.06)

5 Conclusion

Variable selection using the $L_{1,2}$-norm is particularly advantageous in scenarios involving correlated features within the same group. Previous studies have explored various optimizers to address the challenges posed by the composite penalty. In this manuscript, we propose an alternative optimization algorithm tailored to the Exclusive Lasso problem. By approximating the composite penalty norm to be differentiable, our method facilitates the use of Newton-type algorithms, which can outperform coordinate descent methods in terms of computational efficiency.

Our approach employs a straightforward Newton–Raphson algorithm to optimize the objective function. For high-dimensional problems where Hessian computation becomes impractical, we advocate for the use of Quasi-Newton methods, which approximate the Hessian matrix using gradient information from successive iterations. Another viable alternative is the Penalized Iteratively Reweighted Least Squares (PIRLS) method, as suggested by Oelker and Tutz (2017). These techniques have the potential to reduce computational time significantly.

Although direct runtime comparisons with the traditional Exclusive Lasso implemented in the `Exclusive Lasso` R package were not performed, our simulation studies demonstrate the superior performance of our model. Unlike the simulations in Campbell and Allen (2017), which use a fixed group allocation approach assuming known informative variables within each group, our method excels in scenarios with random group allocations, reflecting real-world conditions where such information is typically unavailable. Exclusive Lasso uses the $L_{1,2}$-norm to facilitate the representation of each group. However, our estimation scheme addresses scenarios where a group is non-informative by completely disregarding the group, despite the L_2-norm, by introducing a small threshold. This is why our approach performs better in scenarios with random group allocations.

We applied our model to two real datasets to further validate its efficacy. In the first case, our model successfully identified genes across all modules, which subsequently enabled efficient classification of cancer-sensitive and resistant cohorts. While we did not delve into the biological significance of the selected genes, our findings underscore the importance of selecting variables from diverse modules or pathways. This principle is particularly crucial in multi-omics studies, where disease contributions often involve various omic types, necessitating comprehensive variable selection.

In the second dataset, our model achieved minimal loss with a substantially reduced number of variables compared to the Lasso method, which required more variables to reach the same loss. This efficiency is crucial in fields like portfolio management, where diversification across sectors is essential to mitigate risks. Here, the Exclusive Lasso's ability to select variables from all sectors enhances predictive accuracy for daily returns.

Despite its advantages, the Exclusive Lasso sometimes fails to identify the correct features, especially with random group allocations. To address this, some form of stability selection is already applied to this type of Lasso, as noted by Sun et al. (2020). Incorporating additional stability selection techniques, as proposed

by Meinshausen and Bühlmann (2010), could further enhance the robustness of the method. While this work does not explore such extensions, future work could consider integrating stability selection to improve the Exclusive Lasso's performance further. We also aim to investigate different smoothing strategies for our composite penalty, such as transforming the problem into a quadratic objective function with linear constraints, which potentially then can be tackled by quadratic programming methods. We think of investigating the potential and relevance of Least Absolute Deviation (LAD) regression (Steiger 1983) and its formulation as a linear programming problem in our Exclusive Lasso setup. Finally, we also plan to explore the extension of this approach to Elastic Net penalization, as it could be particularly useful in scenarios involving highly correlated features.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s00180-025-01630-5.

Acknowledgements We sincerely thank the Editor-in-Chief and anonymous reviewers for their insightful comments and valuable suggestions, which significantly contributed to improving the quality of this paper.

Funding Open Access funding enabled and organized by Projekt DEAL. This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project R2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation-Project Number 427806116).

Declarations

Conflict of interest The authors have no Conflict of interest to declare that are relevant to the content of this article.

References

Armijo L (1966) Minimization of functions having Lipschitz continuous first partial derivatives. Pac J Math 16(1):1–3

Belotte J, Fletcher NM, Saed MG, Abusamaan MS, Dyson G, Diamond MP, Saed GM (2015) A single nucleotide polymorphism in catalase is strongly associated with ovarian cancer survival. PLoS ONE 10(8):0135739

Campbell F, Allen GI (2017) Within group variable selection through the exclusive lasso. Electron J Stat 11:4220–4257

Chen C, Mangasarian OL (1996) A class of smoothing functions for nonlinear and mixed complementarity problems. Comput Optim Appl 5(2):97–138

Eddelbuettel D, Balamuta JJ (2018) Extending R with C++: a brief introduction to Rcpp. Am Stat 72(1):28–36. https://doi.org/10.1080/00031305.2017.1375990

Friedman J, Tibshirani R, Hastie T (2010) Regularization paths for generalized linear models via coordinate descent. J Stat Softw 33(1):1–22. https://doi.org/10.18637/jss.v033.i01
Güçkıran K, Cantürk İ, Özyılmaz L (2019) DNA microarray gene expression data classification using SVM, MLP, and RF with feature selection methods relief and LASSO. Süleyman Demirel Üniversitesi Fen Bilimleri Enstitüsü Dergisi 23(1):126–132
Gregoratti D, Mestre X, Buelga C (2021) Exclusive Group Lasso for structured variable selection. arXiv preprint arXiv:2108.10284
Groll A, Schauberger G, Tutz G (2015) Prediction of major international soccer tournaments based on team-specific regularized Poisson regression: an application to the FIFA world cup 2014. J Quant Anal Sports 11(2):97–115
Hitomi K, Kagihara M (2001) Calculation method for nonlinear dynamic least-absolute deviations estimator. J Jpn Stat Soc 31(1):39–51
Huang Y, Liu J (2018) Exclusive sparsity norm minimization with random groups via cone projection. IEEE Trans Neural Netw Learn Syst 29(12):6145–6153
Huang P, Zhang S, Li M, Wang J, Ma C, Wang B, Lv X (2020) Classification of cervical biopsy images based on LASSO and EL-SVM. IEEE Access 8:24219–24228
Kecman V (2005) In: Wang L (ed) Support vector machines: an introduction, pp. 1–47. Springer, Berlin, Heidelberg . https://doi.org/10.1007/10984697_1
Kong D, Fujimaki R, Liu J, Nie F, Ding C (2014) Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. Adv Neural Inf Process Syst 27
Koti M, Gooding RJ, Nuin P, Haslehurst A, Crane C, Weberpals J, Childs T, Bryson P, Dharsee M, Evans K et al (2013) Identification of the IGF1/PI3K/NF κ B/ERK gene signalling networks associated with chemotherapy resistance and treatment response in high-grade serous epithelial ovarian cancer. BMC Cancer 13:1–11
Liu J, Ji S, Ye J (2012) Multi-task feature learning via efficient l2, 1-norm minimization. arXiv preprint arXiv:1205.2631
Lee Y-J, Mangasarian OL (2001) SSVM: a smooth support vector machine for classification. Comput Optim Appl 20:5–22
Lin M, Yuan Y, Sun D, Toh K-C (2020) Adaptive sieving with PPDNA: generating solution paths of exclusive lasso models. arXiv preprint arXiv:2009.08719
Meinshausen N, Bühlmann P (2010) Stability selection. J R Stat Soc Ser B Stat Methodol 72(4):417–473
Mainsiouw L, Ryan ME, Hafizi S, Fleming JC (2023) The molecular and clinical role of Tensin 1/2/3 in cancer. J Cell Mol Med 27(13):1763–1774
Nkansah H, Benyah F, Amankwah H (2021) Smoothing approximations for least squares minimization with l1-norm regularization functional. Int J Anal Appl 19(2):264–279
Oelker M-R, Tutz G (2017) A uniform framework for the combination of penalties in generalized structured models. Adv Data Anal Classif 11(1):97–120
Steiger PBW et al (1983) Least absolute deviations. Applications, and algorithms. Birkhäuser, Boston-Basel-Stuttgart, Theory
Sun Y, Chain B, Kaski S, Shawe-Taylor J (2020) Correlated feature selection with extended exclusive group lasso. arXiv preprint arXiv:2002.12460
Simon N, Friedman J, Hastie T, Tibshirani R (2019) SGL: fit a GLM (or Cox Model) with a combination of Lasso and Group Lasso regularization. R package version 1.3. https://CRAN.R-project.org/package=SGL
Schmidt M, Fung G, Rosales R (2007) Fast optimization methods for l1 regularization: a comparative study and two new approaches. In: Machine learning: ECML 2007: 18th European conference on machine learning, Warsaw, Poland, September 17–21, 2007. Proceedings 18, pp. 286–297. Springer
Tibshirani R (1996) Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol 58(1):267–288
Trinh XB, Tjalma WA, Dirix LY, Vermeulen PB, Peeters DJ, Bachvarov D, Plante M, Berns EM, Helleman J, Van Laere SJ et al (2011) Microarray-based oncogenic pathway profiling in advanced serous papillary ovarian carcinoma. PLoS ONE 6(7):22469
Van Rijsbergen CJ (1979) Information retrieval. 2nd. newton, ma. USA: Butterworth-Heinemann
Wang Y, Li X, Ruiz R (2018) Weighted general group Lasso for gene selection in cancer classification. IEEE Trans Cybern 49(8):2860–2873
Xu Z, Li X, Li H, Nie C, Liu W, Li S, Liu Z, Wang W, Wang J (2020) Suppression of DDX39B sensitizes ovarian cancer cells to DNA-damaging chemotherapeutic agents via destabilizing BRCA1 mRNA. Oncogene 39(47):7051–7062

Yamada M, Koh T, Iwata T, Shawe-Taylor J, Kaski S (2017) Localized Lasso for high-dimensional regression. Artif Intell Stat pp 325–333. PMLR

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol 68(1):49–67

Zhang T, Ghanem B, Liu S, Xu C, Ahuja N (2015) Robust visual tracking via exclusive context modeling. IEEE Trans Cybern 46(1):51–63

Zhang B, Horvath S (2005) A general framework for weighted gene co-expression network analysis. Stat Appl Genet Mol Biol 4(1)

Zhou Y, Jin R, Hoi SC-H (2010) Exclusive Lasso for multi-task feature selection. In: Proceedings of the thirteenth international conference on artificial intelligence and statistics, pp 988–995. JMLR workshop and conference proceedings

Zheng S, Liu W (2011) An experimental comparison of gene selection by lasso and dantzig selector for cancer classification. Comput Biol Med 41(11):1033–1040

Zhao P, Yu B (2006) On model selection consistency of Lasso. J Mach Learn Res 7:2541–2563

Supplementary material: A Newton-based variant of Exclusive Lasso for improved sparse solutions

Dayasri Ravi[1*] and Andreas Groll[1]

[1]Department of Statistics, TU Dortmund University, Vogelpothsweg 87, Dortmund, 44221, NRW, Germany.

*Corresponding author(s). E-mail(s): ravi@statistik.tu-dortmund.de;
Contributing authors: groll@statistik.tu-dortmund.de;

Weighted gene co-expression networks

Weighted gene co-expression networks (WGCNA) are widely used in systems biology to investigate gene functionality on a system-wide scale. We used this approach to group genes into modules based on high topological overlap, where genes with highly correlated expression profiles across samples were clustered together. We considered these modules identified by WGCNA as groups for our Lasso problems. The methodology began with preprocessing the cancer gene expression data to ensure it was clean and normalized for analysis. The dataset was then divided into two parts: one consisting of tumor-sensitive samples and the other of normal samples. We used the tumor data to construct the gene co-expression network. First, we determined the appropriate soft-thresholding power for the network topology. Using a range of powers (1-50), we applied the `pickSoftThreshold` function from the `WGCNA` R package (Langfelder and Horvath, 2008). This function analyzed the scale-free topology fit for different powers. We then plotted the scale-free topology model fit and mean connectivity against the power values. From these plots, we identified a suitable soft-thresholding power.

With the chosen power, we constructed the weighted gene co-expression network using the `blockwiseModules` function from `WGCNA` package. This function identified modules by clustering genes based on their topological overlap. We set an appropriate maximum block size, used a signed network type, and merged modules with a cut height of 0.25. The analysis yielded module **eigengenes**, which represented the first principal component of each module, summarizing the expression profiles of the genes within the module.

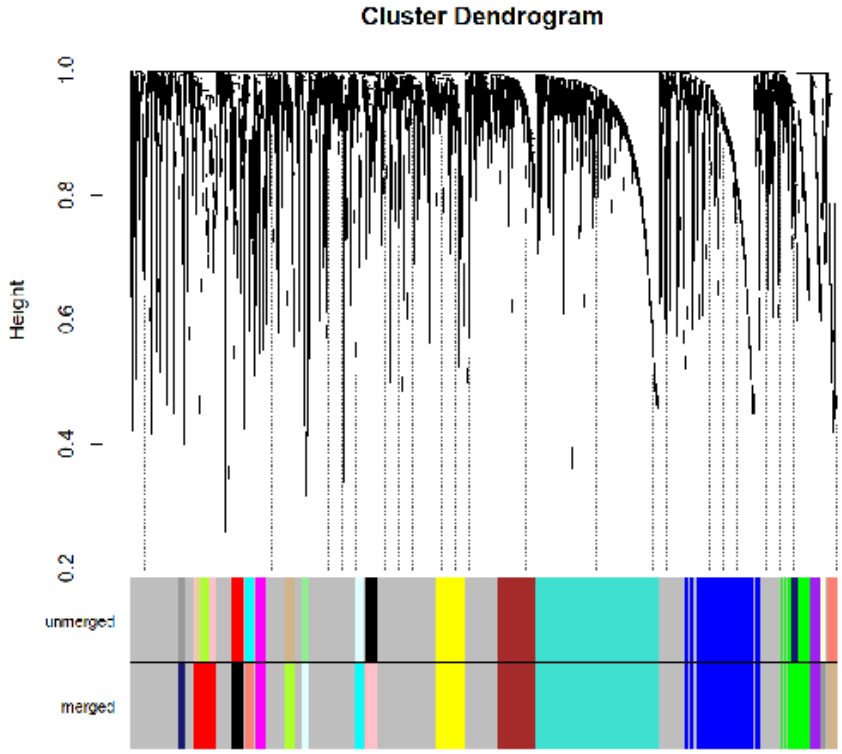

Figure S1: Identification of network modules in OC dataset 1

We examined the number of genes in each module and visualized the dendrogram of gene clustering alongside the module colors, both before and after merging. Finally, these identified modules were used to categorize genes into groups for both tumor and normal samples.

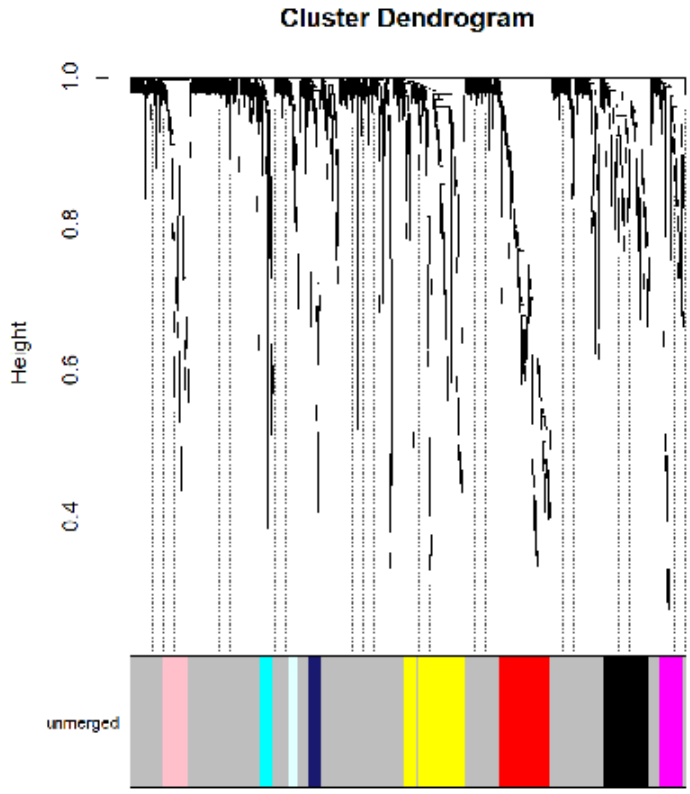

Figure S2: Identification of network modules in OC dataset 2

Figures S1 and S2 show the Cluster Dendrogram of OC dataset 1 and OC dataset 2, respectively. We chose a maximum block size of 3,000 for our analysis. However, because our dataset contained more than 3,000 genes, the figure illustrates the results for only one block as an example.

We excluded unmerged genes labeled as grey that do not belong to any module. The number of genes in each group is shown in Tables S1 and S2 for OC datasets 1 and 2, respectively.

References

Langfelder, P., Horvath, S.: Wgcna: an r package for weighted correlation network analysis. BMC bioinformatics **9**, 1–13 (2008)

Module	Number of Genes
black	62
blue	335
brown	181
cyan	33
green	130
greenyellow	47
lightcyan	28
magenta	51
midnightblue	29
pink	56
purple	51
red	106
salmon	42
tan	45
turquoise	569
yellow	134

Table S1: OC dataset 1: Number of genes in each module

Module	Number of Genes
black	112
blue	450
brown	200
cyan	31
green	135
greenyellow	55
lightcyan	26
lightgreen	24
lightyellow	24
magenta	61
midnightblue	31
pink	66
purple	58
red	129
salmon	44
tan	47
turquoise	3894
yellow	144

Table S2: OC dataset 2: Number of genes in each module

Article V

Time-to-event prediction for grouped variables using Exclusive Lasso

Dayasri Ravi, Andreas Groll

Department of Statistics, TU Dortmund University

Abstract

The integration of high-dimensional genomic data and clinical data into time-to-event prediction models has gained significant attention due to the growing availability of these datasets. Traditionally, a Cox regression model is employed, concatenating various covariate types linearly. Given that much of the data may be redundant or irrelevant, feature selection through penalization is often desirable. A notable characteristic of these datasets is their organization into blocks of distinct data types, such as methylation and clinical predictors, which requires selecting a subset of covariates from each group due to high intra-group correlations. For this reason, we propose utilizing Exclusive Lasso regularization in place of standard Lasso penalization. We apply our methodology to a real-life cancer dataset, demonstrating enhanced survival prediction performance compared to the conventional Cox regression model.

1 Introduction

In recent years, advancements in high-throughput genomic technologies have led to the availability of high-dimensional datasets, including DNA methylation, mRNA expression, and copy number variation, in addition to traditional clinical variables. These datasets may provide valuable information on the mechanisms of a certain disease, prompting the development of various methods to identify influential genomic and clinical characteristics for improved prognostic modeling.

A common objective in clinical research is the prediction of patient survival outcomes. The Cox proportional hazards (PH) model (Cox, 1972) is widely used for this purpose, as it not only facilitates survival prediction but also enables the assessment of the impact of predictor variables on survival. However, given the high-dimensional nature of genomic datasets, variable selection becomes a critical step in model construction. To address this, an L_1-penalized Cox model, such as the Lasso, is often employed to identify the most relevant features in time-to-event modeling (Tibshirani, 1997).

Despite its effectiveness, this approach presents several limitations. First, standard Lasso-based methods do not inherently account for grouped variables, which is particularly relevant in genomic studies where genes are often organized in biological pathways.

Ignoring such group structures may lead to suboptimal feature selection and loss of biologically meaningful information. Additionally, large sets of genomic features often overshadow low-dimensional clinical variables, such as tumor size and nodal status. This is a significant drawback, as clinicopathologic variables have been demonstrated to play a crucial role in oncological studies, and predictive performance improves when both clinical and genomic data are integrated (Ma et al., 2007; Herrmann et al., 2021). Finally, Lasso-based selection methods have been shown to produce a relatively high rate of false positives in certain settings, which may limit their reliability in time-to-event analysis, depending on the context. (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006).

Several statistical methods have been proposed to incorporate grouped variables in the Cox PH model. Although Elastic Net Cox (Simon et al., 2011b) does not explicitly enforce group selection, it tends to select correlated variables together, unlike pure Lasso, which typically selects only one from a correlated set. This behavior results from its combination of L_1 and L_2-norm penalties. However, it does not guarantee that entire groups of variables will be retained or removed together. One of the most common approaches is Group Lasso (Kim et al., 2012), which applies $L_{2,1}$-norm regularization. This method enforces sparsity across groups using the L_1-norm while applying the L_2-norm within each group to regularize coefficients together. However, Group Lasso performs poorly when dealing with highly correlated groups, such as those found in multi-omics datasets. In such cases, it tends to select variables from only a few dominant groups, often overlooking smaller or lower-dimensional groups like clinical variables, which are essential for time-to-event prediction. This requires a method that can ensure the selection of variables for each group rather than selecting entire groups of variables. To overcome this challenge, Sparse Group Lasso was introduced (Simon et al., 2013), incorporating sparsity at both the group and individual variable levels. Another advancement in this area is the Integrative L_1-Penalized Regression with Penalty Factors (IPF-Lasso; Boulesteix et al., 2017), which allows for different penalty terms across variable groups, either based on prior knowledge or data-driven selection. However, none of these methods guarantee the selection of at least one variable from every group, which can lead to the exclusion of smaller yet important groups from the model. In biomedical studies, representing all relevant groups is important for gaining a complete understanding of underlying relationships.

To overcome this limitation, we propose the use of Exclusive Lasso regularization (Campbell and Allen, 2017). Exclusive Lasso encourages intra-group sparsity through the L_1-norm while promoting inter-group selection via the L_2-norm, ensuring that at least one variable is selected from each group. The properties of this regularization have been well studied in the literature (Campbell and Allen, 2017; Gregoratti et al., 2021).

In our previous work, we demonstrated the superior performance of Exclusive Lasso over traditional Lasso in GLM settings with high within-group correlation (Ravi and Groll, 2025; note that a preliminary compact version of this work can also be found in Ravi and Groll, 2024). However, this approach has not yet been transferred to time-to-event prediction. In this study, we extend Exclusive Lasso to the Cox PH model, introducing it as a practical alternative for selecting informative predictors from different groups. Our goal

is to integrate these selected features into a sparse prediction model while ensuring that no group is overlooked.

We assess the performance of our proposed method by comparing it to traditional approaches that account for grouping effects, such as Elastic Net Cox, Sparse Group Lasso, and IPF-Lasso. Through extensive simulation studies, we show that our method consistently outperforms these alternatives across a range of scenarios. Additionally, we evaluate the practical applicability of our model by using it for survival prediction in a bladder cancer study. In addition to the standard prediction errors, we compare the biomarkers selected by each model and highlight the importance of Exclusive Lasso in selecting clinical and low-dimensional variables that other models fail to capture.

The remainder of the manuscript is structured as follows. Section 2 introduces the Exclusive Lasso problem in the Cox PH framework. In Section 3, we present the simulation scenarios and compare our method with other Lasso procedures. The applicability of our model is demonstrated in Section 4 using the aforementioned application example. Finally, Section 5 concludes.

2 Methods

In this section, we first briefly review traditional methods for handling grouped predictors within the Cox PH framework and then introduce the Exclusive Lasso regularization in the Cox PH model.

Let $i = 1, \ldots, n$ denote the observations (patients) in the cohort. For each patient, we observe $(t_i, \delta_i, \mathbf{x}_i)$, where t_i is the event or censoring time for patient i, δ_i is the censoring indicator, which is 1 if an event is observed and 0 if the observation is censored and $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})^T$ is the vector of covariates associated with patient i.

The partial log-likelihood function for the Cox PH model (Cox, 1972), with parameter vector $\boldsymbol{\beta}$, is given by:

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \delta_i \left[\mathbf{x}_i^T \boldsymbol{\beta} - \log \left(\sum_{l \in R(t_i)} \exp(\mathbf{x}_l^T \boldsymbol{\beta}) \right) \right] , \tag{2.1}$$

where $R(t_i)$ represents the risk set at time t_i, which includes all individuals who are still at risk (i.e., uncensored and have not yet experienced the event) at the time of observation.

In high-dimensional scenarios, where the number of covariates p by far exceeds the number of patients n, the estimation of the coefficients is typically performed by introducing a penalty term, $P(\boldsymbol{\beta})$, to the partial log-likelihood function. The estimation is then carried out by maximizing the penalized partial log-likelihood function, given by

$$\ell_{\text{pen}}(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) - \lambda P(\boldsymbol{\beta}) , \tag{2.2}$$

where $\lambda \geq 0$ is the penalty parameter. The most common penalty term is the L_1-norm penalty, i.e., $P(\boldsymbol{\beta}) = \sum_{k=1}^{p} |\beta_k|$. This penalty shrinks the coefficients towards zero and

forces some of the coefficients corresponding to less important variables to be exactly zero, effectively performing variable selection.

We focus on variable selection in scenarios where the predictors are divided into predefined, disjoint groups. For instance, in the context of multi-omics data, the variables may include different types, such as genomics, epigenomics, and transcriptomics, in addition to clinical and pathological data. We assume that the indices of the true parameter vector $\boldsymbol{\beta}$ are divided into non-overlapping groups. Let $\mathcal{G}$ be a collection of non-overlapping predefined groups $\{1, \ldots, G\}$ of indices g such that the union of all groups covers the entire set of indices, i.e.,

$$\bigcup_{g \in \mathcal{G}} = \{1, \ldots, p\} .$$

Elastic Net

Elastic Net (Zou and Hastie, 2005) is a regularization method that combines the L_1 (Lasso) and L_2 (Ridge) penalties. The L_1 penalty encourages sparsity by shrinking some coefficients to zero, while the L_2 penalty promotes the inclusion of correlated variables in groups. This combination allows groups of correlated variables to be selected together. Elastic Net is particularly effective in situations where predictors are highly correlated within groups. The Elastic Net penalty is defined as :

$$P(\boldsymbol{\beta}) = \alpha \sum_{k=1}^{p} |\beta_k| + \frac{1}{2}(1-\alpha) \sum_{k=1}^{p} \beta_k^2 ,$$

where $\alpha \in (0, 1)$ is the mixing parameter that controls the balance between the L_1 and L_2 penalties.

Sparse Group Lasso

Sparse Group Lasso (Simon et al., 2013) is another method that uses a combination of L_1- and L_2-norm penalties to encourage sparsity both across groups and within each group. The penalty term for Sparse Group Lasso is given by:

$$P(\boldsymbol{\beta}) = (1-\alpha) \sum_{g \in \mathcal{G}} \sqrt{p_g} \|\boldsymbol{\beta}_g\|_2 + \alpha \sum_{k=1}^{p} |\beta_k| ,$$

where $\boldsymbol{\beta}_g$ represents the coefficients in group $g \in \mathcal{G}$, and p_g is the number of variables in group g. This promotes group-wise selection, where all variables in a group are either included or excluded together. The second term is the Lasso penalty applied to individual coefficients, promoting sparsity at the level of individual predictors. When the parameter $\alpha = 0$, the Sparse Group Lasso reduces to the standard Group Lasso, and when $\alpha = 1$, it becomes the Lasso.

Integrative Lasso with penalty factors (IPF-Lasso)

The Integrative Lasso with penalty factors (IPF-Lasso; Boulesteix et al., 2017) was introduced for prediction based on multi-omics datasets where there are several modalities (groups) of variables. The main idea of IPF-Lasso is to apply Lasso to each group and introduce penalty factors for different groups of variables, which can be selected according to the desired weighting of the groups or by cross-validation (CV). The IPF-Lasso penalty is defined as

$$\sum_{g \in \mathcal{G}} \lambda_g \|\boldsymbol{\beta}_g\|_1 \ ,$$

where λ_g is the penalty factor applied to the variables in group g. These penalty factors are chosen by CV via a grid search over a list of prespecified candidate vectors. However, this can be a time-consuming process. To avoid manually defining candidate vectors, we follow the procedure outlined in the Two-step IPF-Lasso (Schulze, 2017). In Step 1 of the process, before applying IPF-Lasso, a standard Lasso or Ridge regression is performed, and the arithmetic mean of the estimated coefficients can be considered as potential penalty parameters.

Exclusive Lasso

The Exclusive Lasso (Campbell and Allen, 2017) enforces structured sparsity by ensuring that at least one variable is selected from each predefined group. It combines L_1- and L_2-norm penalties, where the L_1 penalty within each group promotes the selection of informative variables, while the L_2-norm across groups prevents entire groups of coefficients from being set to zero. This approach ensures that even low-dimensional groups are represented while selecting only the most relevant variables from high-dimensional groups. The Exclusive Lasso penalty, which can be added to the Cox PH partial log-likelihood, is defined as:

$$P(\boldsymbol{\beta}) = \frac{1}{2} \sum_{g \in \mathcal{G}} \left(\sum_{k \in g} |\beta_k| \right)^2 . \tag{2.3}$$

The composite nature of the penalty term makes the estimation of the Exclusive Lasso problem challenging. Several strategies have been developed to tackle this challenge. One approach utilizes proximal point algorithms based on dual Newton methods (Lin et al., 2020), while others employ iterative re-weighted techniques to refine the estimation process (Kong et al., 2014; Sun et al., 2020). An alternative strategy reformulates the problem in a Lasso framework and applies a bisection algorithm, taking advantage of Lasso's piecewise linear properties (Sun et al., 2020).

More recently, a fast optimization method leveraging the iterative shrinkage-thresholding algorithm (FISTA) has been proposed to improve computational efficiency (Huang and Liu, 2018). Another approach transforms the penalty into a differentiable one by applying a simple quadratic approximation, allowing it to be efficiently solved using a Newton-based algorithm (Ravi and Groll, 2025). However, to the best of our knowledge, none of these methods have been extended to time-to-event prediction yet.

To address this challenge, we employ the coordinate descent method with soft-thresholding. As highlighted by Campbell and Allen (2017), the Exclusive Lasso penalty cannot be expressed as a sum of separable functions, i.e.,

$$P(\boldsymbol{\beta}) \neq \sum_{j=1}^{p} P_j(\beta_j).$$

This implies that a simultaneous update of all variables is not feasible. Instead, estimation is performed using a coordinate descent algorithm, where each variable is updated sequentially while keeping the others fixed.

Our approach adopts a coordinate descent framework tailored for Group Lasso regularization (Yuan and Lin, 2006), which has been shown to be efficient in solving high-dimensional problems. The proposed algorithm is presented in Algorithm 1.

The gradient component for covariate j in the Cox PH model from Equation (2.1) is defined as:

$$\hat{r}_j = \sum_{i=1}^{n} \delta_i \left[x_{ij} - \frac{\sum\limits_{l \in R(t_i)} x_{lj} \exp(\mathbf{x}_l^T \boldsymbol{\beta})}{\sum\limits_{l \in R(t_i)} \exp(\mathbf{x}_l^T \boldsymbol{\beta})} \right],$$

where x_{lj} represents the observed value of covariate j for individual l, and $R(t_i)$ denotes the risk set at time t_i. This formulation ensures that only individuals for whom $\delta_i = 1$ (i.e., those who experience an event) contribute to the estimation of β_j.

We apply coordinate descent to maximize the penalized partial log-likelihood defined in Equation (2.2), using the soft-thresholding operator $S(z, \lambda)$ given by:

$$S(z, \lambda) = \text{sign}(z) \max(|z| - \lambda, 0). \tag{2.4}$$

The soft-thresholding operator shrinks coefficients toward zero by subtracting a threshold λ, setting them exactly to zero when their magnitude falls below this threshold. This encourages sparsity and facilitates automatic variable selection.

Since we update one coefficient β_j at a time while holding others fixed, we use $g \setminus j$ to denote the set of indices in group g excluding index j. The corresponding penalty term for β_j is then updated as:

$$\tilde{P}_j = \lambda \sum_{l \in g \setminus j} |\beta_l|.$$

This penalty encourages competition among variables within the same group, allowing only a few features to get selected. It promotes sparsity by shrinking coefficients, especially when the penalty is large. As a result, β_j is pushed toward zero when other covariates in the group have large values, reducing redundancy among correlated variables.

Furthermore, we refer readers to Theorem 4 of Campbell and Allen (2017), which provides proof that the Exclusive Lasso coordinate descent algorithm converges to the global minimum.

Algorithm 1 Exclusive Lasso coordinate descent for Cox PH model

Require: Initial coefficients $\boldsymbol{\beta}^0 \in \mathbb{R}^p$, tolerance $\epsilon > 0$, regularization parameter $\lambda > 0$, group structure $\mathcal{G}$, design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, event times $\mathbf{T} \in \mathbb{R}^n$, event indicators $\boldsymbol{\delta} \in \{0,1\}^n$

Precompute: Sort the data in increasing order of event times $\mathbf{T}$
while $||\boldsymbol{\beta}^{(k+1)} - \boldsymbol{\beta}^{(k)}|| > \epsilon$ **do**
  for all groups $g \in \mathcal{G}$ **do**
    for all features $j \in g$ **do**
      Compute gradient component for covariate j :

$$\hat{r}_j = \sum_{i=1}^{n} \delta_i \left[x_{ij} - \frac{\sum\limits_{l \in R(t_i)} x_{lj} \exp(\mathbf{x}_l^T \boldsymbol{\beta})}{\sum\limits_{l \in R(t_i)} \exp(\mathbf{x}_l^T \boldsymbol{\beta})} \right],$$

      Compute Exclusive Lasso Penalty terms:

$$\tilde{P}_j = \lambda \sum_{l \in g \setminus \{j\}} |\beta_l|$$

      Compute Hessian approximation:

$$H_j = \sum_{i=1}^{n} \delta_i x_{ij}^2$$

      Update β_j using soft-thresholding from Equation (2.4):

$$\beta_j^{(k+1)} = S\left(\frac{\hat{r}_j}{H_j + \lambda}, \frac{\tilde{P}_j}{H_j + \lambda} \right)$$

    end for
  end for
end while
return $\hat{\boldsymbol{\beta}}$

Figure 1 displays the regularization paths for both Exclusive Lasso (left) and Lasso (right). The variables are divided into five distinct groups, with each group containing exactly one signal variable and the rest being noise. The signal variables are highlighted using different colors to distinguish their respective groups. Exclusive Lasso encourages within-group sparsity, driving most coefficients to zero while retaining only one active variable per group. As a result, it maintains exactly five active variables, one from each group, even at large values of λ. In contrast, Lasso applies shrinkage without regard to group structure and may eliminate informative variables or retain multiple variables from the same group.

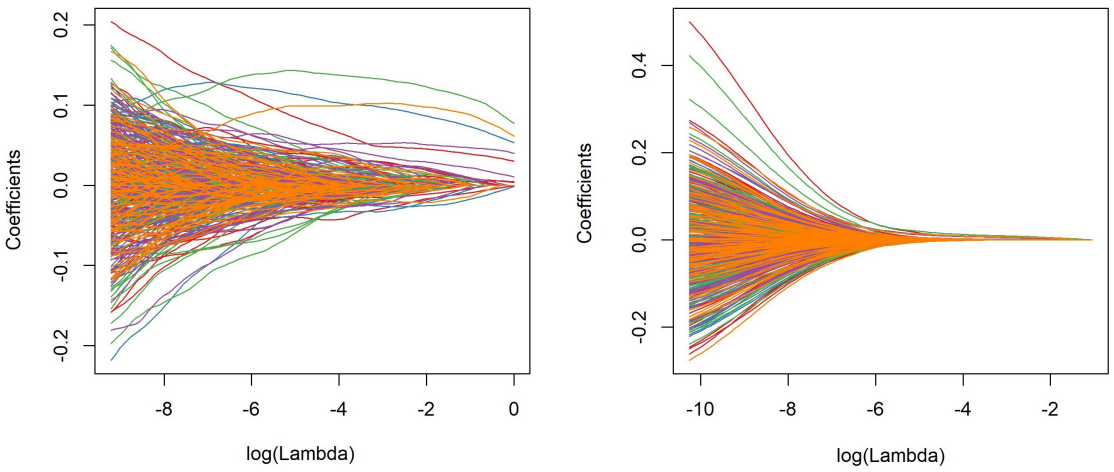

Figure 1: Regularization paths for Exclusive Lasso (*left*) and Lasso (*right*) from a simulation study, where variables are evenly distributed into five groups (shown in distinct colors), with each group containing one true signal variable. In the Exclusive Lasso model, the true variables remain active unless all other variables in their group shrink to zero. In contrast, the Lasso model selects variables without considering the group structure, allowing multiple variables from the same group to be included.

3 Simulations

In this section, we present a detailed simulation study to evaluate the performance of our method across different scenarios.

3.1 Setting

We simulate $n = 500$ observations and $p = 500$ variables from a multivariate Gaussian distribution with a Toeplitz covariance matrix $\mathbf{\Sigma}$, where the entries $\Sigma_{i,j} = 0.6^{|i-j|}$ for variables in the same group, and $\Sigma_{i,j} = 0.3^{|i-j|}$ for variables in different groups. Altogether, we use a moderate correlation, resulting in a high correlation within groups and a low correlation between groups. Event times are simulated using a Cox PH model framework, where the hazard function depends on a baseline hazard and a linear combination of the predictors. We consider a baseline median event time of eight years. The true coefficients are drawn from a uniform distribution between 0.5 and 1.5. Independent censoring times are simulated using an exponential distribution with a rate of 0.02, assuming constant censoring hazard over time.

We assume that the variables are divided into five groups and consider three different simulation scenarios for grouping them. The total number of signal variables in all three scenarios is set to 5, 10, and 20, respectively.

Table 1 provides a detailed overview of the grouping structure used across all three scenarios. In the first scenario, we allocate an equal number of variables to each group, with each group containing 100 variables. This represents an ideal setting, as Exclusive Lasso is expected to perform well when at least one signal variable is present in each group.

In the second and third scenarios, we introduce unequal group sizes. In Scenario 2, the variables are distributed as (15, 20, 85, 180, 200), while in Scenario 3, the distribution is (5, 295, 10, 90, 100). Additionally, the signal variables are also unequally distributed among the groups.

Scenario 1				Scenario 2				Scenario 3			
Variables per group	True variables per group			Variables per group	True variables per group			Variables per group	True variables per group		
100	1	2	4	15	1	1	2	5	1	1	2
100	1	2	4	20	1	2	2	295	1	2	6
100	1	2	4	85	1	1	1	10	1	1	4
100	1	2	4	180	1	4	10	90	1	2	6
100	1	2	4	200	1	2	5	100	1	4	2

Table 1: Description of grouping structure of both signal and noise variables across three simulation scenarios.

We simulate an independent validation dataset consisting of $n = 500$ observations to evaluate model performance. The penalty parameter λ is selected via 5-fold CV for all models. For the IPF-Lasso, we choose the value of λ by maximizing the cross-validated predictive log-likelihood via a 5-fold CV with 10 repeats. Although we initially aimed to assess our simulations using Sparse Group Lasso with a cross-validated mixing parameter α, this approach significantly increased memory and time requirements. Thus, we fix $\alpha = 0$ to maintain computational efficiency comparable to other models. This choice corresponds to the standard Group Lasso and avoids using the default $\alpha = 0.95$ recommended by the authors of the `SGL R` package (Simon et al., 2019).

We report the results using variable selection accuracy, defined as the proportion of true positives and true negatives among all variables, along with the F1 score, false discovery rate (FDR), and integrated Brier score (IBS). The F1 score (Van Rijsbergen, 1979) is defined as the harmonic mean of precision and recall, accounting for both false positives and false negatives. The metric ranges from 0 to 1, with larger values indicating a better balance between precision and recall. The Brier score (Graf et al., 1999) at a given time point t represents the average squared distances between the observed event status and the predicted survival probability. The IBS provides an overall performance of the model by integrating the Brier score at all available time points. A lower value of IBS is desired as it indicates that the model's predicted probabilities are closer to the true probabilities across all available time points.

3.2 Results

We compare our proposed extension of Exclusive Lasso for Cox PH models with the models described in Section 2. We use the implementations available in the R packages: `glmnet`

No. of true variables	Metric	Elastic Net	Exclusive Lasso	Group Lasso	IPF
5	Selection Accuracy	0.86 (0.001)	**0.99 (0.000)**	0.01 (0.000)	0.93 (0.001)
	F1 score	0.13 (0.001)	**0.67 (0.007)**	0.02 (0.000)	0.23 (0.002)
	False discovery rate	0.93 (0.000)	**0.49 (0.010)**	0.99 (0.000)	0.87 (0.001)
	Integrated Brier score	0.56 (0.001)	**0.42 (0.000)**	0.44 (0.001)	0.53 (0.000)
10	Selection Accuracy	0.83 (0.000)	**0.98 (0.000)**	0.02 (0.000)	0.90 (0.001)
	F1 score	0.19 (0.000)	**0.65 (0.002)**	0.04 (0.000)	0.25 (0.001)
	False discovery rate	0.89 (0.000)	**0.52 (0.003)**	0.98 (0.000)	0.85 (0.001)
	Integrated Brier score	0.61 (0.001)	**0.43 (0.000)**	0.47 (0.000)	0.52 (0.001)
20	Selection Accuracy	0.81 (0.001)	**0.91 (0.000)**	0.04 (0.000)	0.89 (0.001)
	F1 score	0.30 (0.001)	**0.47 (0.001)**	0.08 (0.000)	0.42 (0.001)
	False discovery rate	0.83 (0.001)	**0.69 (0.001)**	0.96 (0.000)	0.74 (0.001)
	Integrated Brier score	0.65 (0.000)	**0.43 (0.000)**	0.49 (0.000)	0.55 (0.000)

Table 2: Average performance metrics (standard errors in brackets) for different models across varying numbers of signal variables in Scenario 1 over 100 iterations; best-performing modeling approach per setting in bold font.

(Friedman et al., 2010; Simon et al., 2011a) for Elastic Net Cox, `SGL` (Simon et al., 2019) for Group Lasso, and `ipflasso` (Boulesteix et al., 2019) for IPF-Lasso. We generate 100 random test datasets and report the average selection accuracy, F1 score, false discovery rate, and IBS for different numbers of true variables across Scenarios 1, 2, and 3 in Tables 2–4, respectively. Figure 2 visually represents the average selection accuracy and false discovery rate over 100 random replications. Exclusive Lasso demonstrates the largest selection accuracy and F1 score across all three simulation scenarios.

For most models, except Group Lasso, accuracy decreases as the number of signal variables increases. Group Lasso, on the other hand, shows a slight improvement with more signal variables, as it tends to select all variables within a group. However, despite this slight improvement, its overall performance remains quite poor.

Exclusive Lasso's performance declines more significantly in Scenarios 2 and 3, where variable allocation is more random (see Table 3). As the number of signal variables increases, its accuracy becomes comparable to that of IPF-Lasso. In contrast, Elastic Net maintains consistent performance across all scenarios. This is because Elastic Net does not explicitly account for grouping, so variations in the number of variables per group do not impact its selection process.

Regarding the false discovery rate (FDR), Exclusive Lasso is the only model where FDR increases as the number of signal variables rises. For the other models, FDR decreases with more signal variables. This occurs because, as the number of variables or groups increases, Exclusive Lasso tends to select variables from every group, even if some groups contain only non-informative variables. This behavior is especially evident in Scenario 3, where signal variables are not evenly distributed among groups (see Table 4). However, despite this increase in FDR, Exclusive Lasso still maintains a lower false discovery rate than the other models.

However, we do not observe significant differences in IBS scores across different scenarios

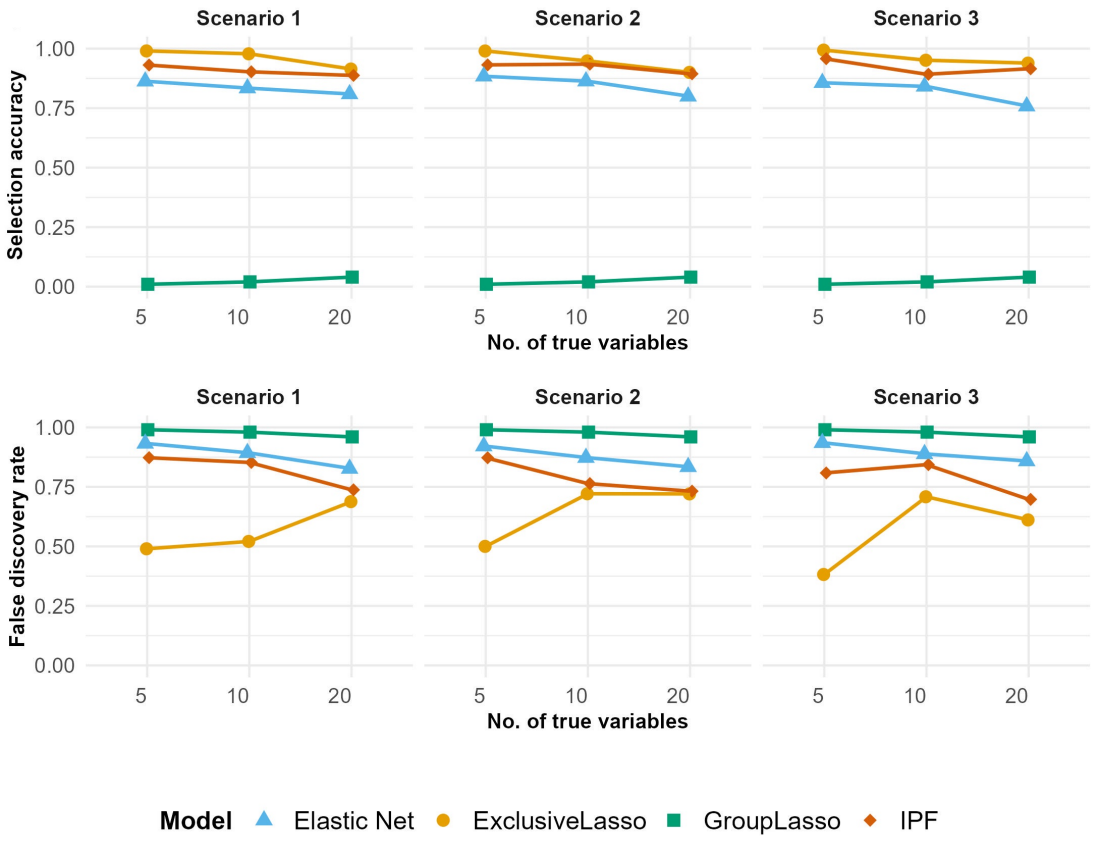

Figure 2: Selection accuracy and false discovery rate across three scenarios for varying numbers of signal variables. The results are averaged over 100 randomly generated datasets.

or varying numbers of true variables. This could be due to the high correlation among variables, allowing the models to produce survival probability estimates that remain close to the true values despite differences in selection accuracy. As a result, IBS scores remain relatively stable across all models. Nonetheless, Exclusive Lasso exhibits slightly better predictive performance compared to the other models.

Overall, we find that Exclusive Lasso outperforms the other models in scenarios where variables are highly correlated and grouped. While its performance declines slightly in randomly allocated scenarios, it still remains superior to the other models. The only model that comes close to Exclusive Lasso is IPF-Lasso, and even then, only when the number of signal variables is large.

No. of true variables	Metric	Elastic Net	Exclusive Lasso	Group Lasso	IPF
5	Selection Accuracy	0.88 (0.000)	**0.99 (0.000)**	0.01 (0.000)	0.93 (0.000)
	F1 score	0.15 (0.000)	**0.67 (0.000)**	0.02 (0.000)	0.23 (0.001)
	False discovery rate	0.92 (0.000)	**0.50 (0.000)**	0.99 (0.000)	0.87 (0.000)
	Integrated Brier score	0.55 (0.001)	**0.43 (0.000)**	0.49 (0.001)	0.53 (0.001)
10	Selection Accuracy	0.86 (0.001)	**0.95 (0.000)**	0.02 (0.000)	0.94 (0.001)
	F1 score	0.23 (0.001)	**0.44 (0.001)**	0.04 (0.000)	0.38 (0.003)
	False discovery rate	0.87 (0.001)	**0.72 (0.001)**	0.98 (0.000)	0.76 (0.002)
	Integrated Brier score	0.62 (0.001)	**0.43 (0.000)**	0.51 (0.001)	0.58 (0.001)
20	Selection Accuracy	0.80 (0.001)	**0.90 (0.000)**	0.04 (0.000)	0.89 (0.000)
	F1 score	0.29 (0.001)	**0.43 (0.000)**	0.08 (0.000)	0.42 (0.000)
	False discovery rate	0.83 (0.000)	**0.72 (0.000)**	0.96 (0.000)	0.73 (0.000)
	Integrated Brier score	0.65 (0.001)	**0.43 (0.000)**	0.51 (0.000)	0.53 (0.000)

Table 3: Average performance metrics (standard errors in brackets) for different models across varying numbers of signal variables in Scenario 2 over 100 iterations; best-performing modeling approach per setting in bold font.

No. of true variables	Metric	Elastic Net	Exclusive Lasso	Group Lasso	IPF
5	Selection Accuracy	0.86 (0.000)	**0.99 (0.000)**	0.01 (0.000)	0.96 (0.000)
	F1 score	0.12 (0.000)	**0.76 (0.007)**	0.02 (0.000)	0.32 (0.001)
	False discovery rate	0.94 (0.000)	**0.38 (0.007)**	0.99 (0.000)	0.81 (0.001)
	Integrated Brier score	0.56 (0.001)	**0.44 (0.000)**	0.46 (0.001)	0.54 (0.001)
10	Selection Accuracy	0.84 (0.001)	**0.95 (0.001)**	0.02 (0.000)	0.89 (0.001)
	F1 score	0.20 (0.001)	**0.45 (0.003)**	0.04 (0.000)	0.27 (0.001)
	False discovery rate	0.89 (0.000)	**0.71 (0.002)**	0.98 (0.000)	0.84 (0.001)
	Integrated Brier score	0.61 (0.001)	**0.43 (0.000)**	0.47 (0.001)	0.61 (0.001)
20	Selection Accuracy	0.76 (0.001)	**0.94 (0.000)**	0.04 (0.000)	0.92 (0.000)
	F1 score	0.25 (0.001)	**0.54 (0.002)**	0.08 (0.000)	0.45 (0.001)
	False discovery rate	0.86 (0.000)	**0.61 (0.002)**	0.96 (0.000)	0.70 (0.001)
	Integrated Brier score	0.68 (0.001)	**0.43 (0.000)**	0.49 (0.000)	0.53 (0.000)

Table 4: Average performance metrics (standard errors in brackets) for different models across varying numbers of signal variables in Scenario 3 over 100 iterations; best-performing modeling approach per setting in bold font.

4 Application

Next, we apply our proposed method to a real-world gene expression dataset. The regularization parameter λ is tuned via CV for all models.

Bladder cancer (BC) is one of the most commonly diagnosed urinary cancers worldwide, with its incidence steadily increasing each year. This rise may be linked to factors such as tobacco use and an aging population. Although the 5-year survival rate for BC is relatively high at 77%, the recurrence rate remains a significant concern. Beyond genetic signatures, numerous risk factors contribute to BC development, including gender, smoking pattern, and occupational exposure to carcinogens (Cumberbatch et al., 2018). Therefore, it is crucial

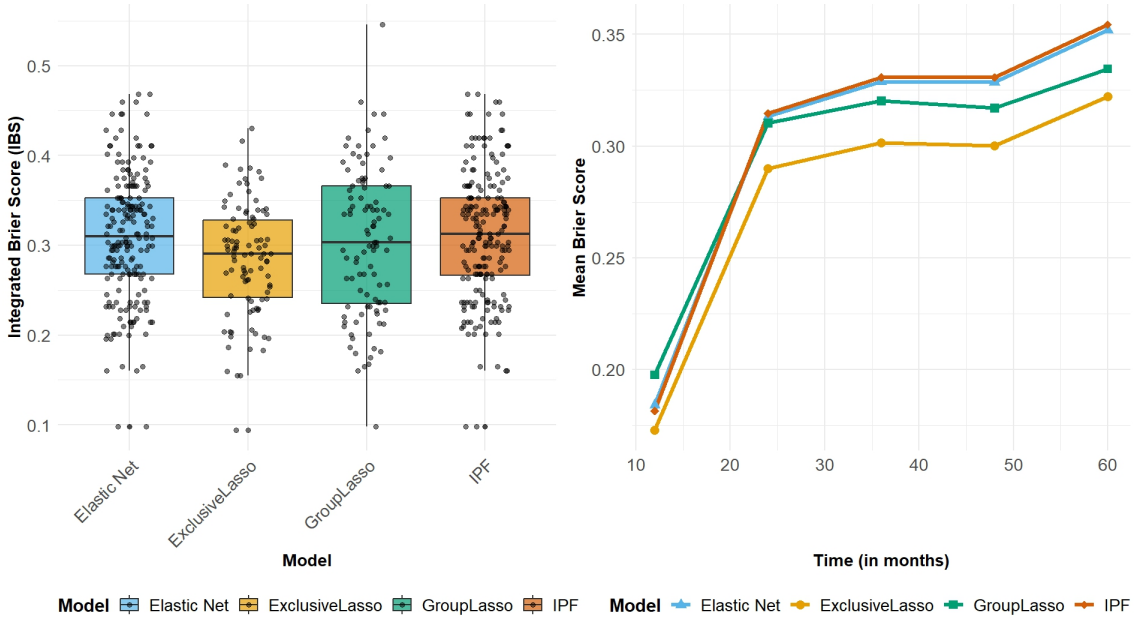

Figure 3: Prediction errors for Bladder gene expression study. *Left*: Boxplots of the integrated Brier score evaluated up to 60 months across 100 random training-test-data splits. *Right*: Mean Brier scores calculated at 10, 20, 30, 40, 50, and 60 months, averaged over 100 random training-test-data splits.

to incorporate both clinical risk factors and sensitive biomarkers when predicting overall survival in bladder cancer patients.

We analyze the BC dataset retrieved from the Gene Expression Omnibus (GEO) database (URL: `https://www.ncbi.nlm.nih.gov/geo/`) using the "GEOquery" `Bioconductor` R package, with the GEO accession GSE31684 (Riester et al., 2012). The dataset includes gene expression data for 54,675 genes from 93 patients. For data preprocessing, we apply a variance filter to select genes with high variance, as previous studies have shown that using a variance filter before fitting a regularized Cox PH regression model can improve the performance of regularized Cox regression models and lead to stable feature selection (Bommert et al., 2022). We categorize the variables into two groups: clinical and gene expression. The clinical group includes age (in years), tumor stage (Ta/T1, T2, T3, T4), nomogram score, and packs smoked per year. The data is split into training (70%) and testing (30%) sets, and this process is repeated for 100 times. For each split, we compute prediction errors on the test set and identify the top 10 most frequently selected variables by each model on the training data. This frequency-based approach is motivated by the principle of stability selection (Meinshausen and Bühlmann, 2010), which aims to identify variables that are consistently selected across multiple resampled datasets, thereby improving the robustness and reliability of variable selection in high-dimensional settings.

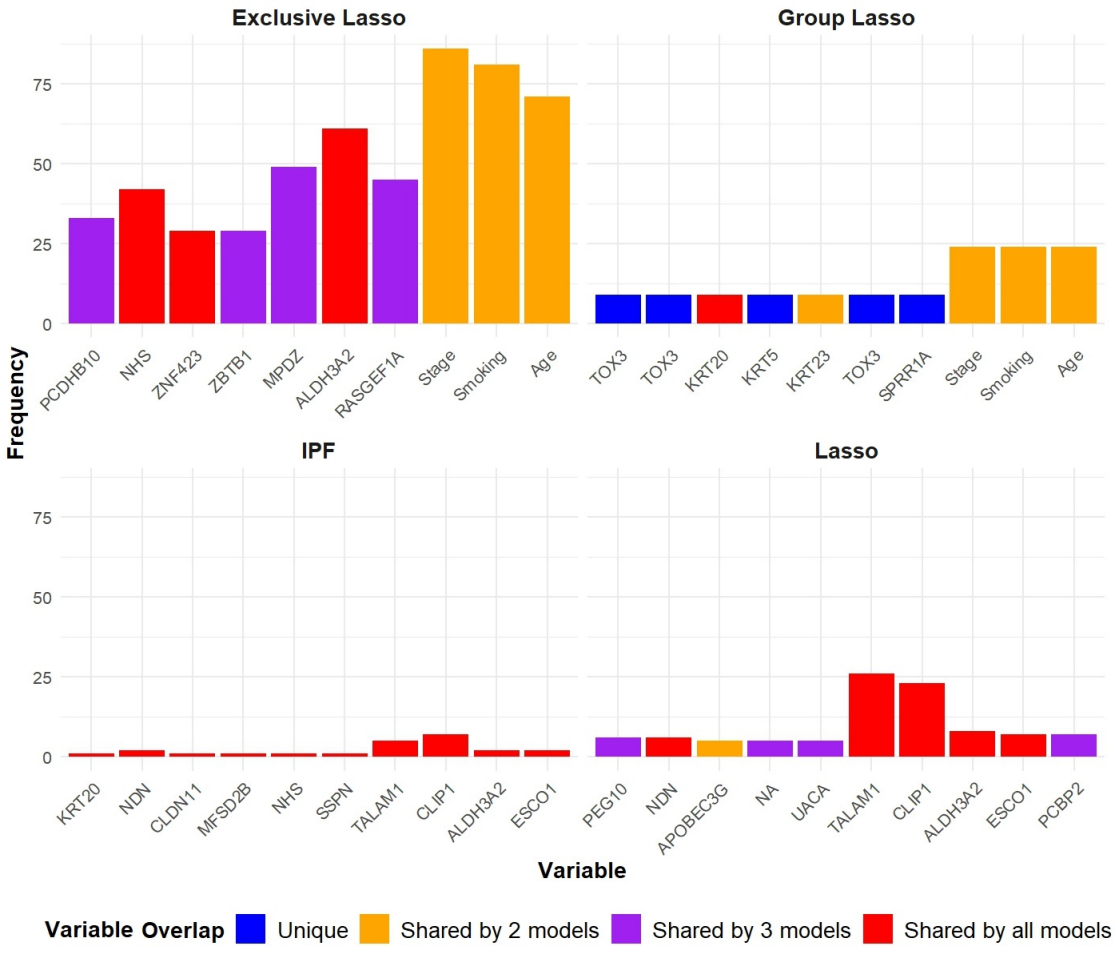

Figure 4: The top 10 most frequently selected variables by the different models on the training set of the Bladder cancer gene expression study for 100 random subsamples.

We report the Brier scores computed up to 5 years for all the models discussed in Section 2. From Figure 3, we observe that Exclusive Lasso yields the lowest IBS curve. Although there is no substantial difference in the IBS across models, the right plot in Figure 3 shows that Exclusive Lasso consistently gives the lowest mean Brier score at each time point. Elastic Net and IPF-Lasso perform almost identically. Group Lasso shows larger Brier scores for events between 10 and 20 months, but beyond that, its Brier score is lower than that of Elastic Net and IPF-Lasso.

Figure 4 displays the top 10 most frequently selected variables by all models. We observe that the clinical variables "Stage", "Smoking", and "Age" are selected more frequently by Exclusive Lasso than by any other model. In 100 iterations, the variable "Stage" is selected more than 80 times. Despite being part of a low-dimensional clinical group, Exclusive Lasso

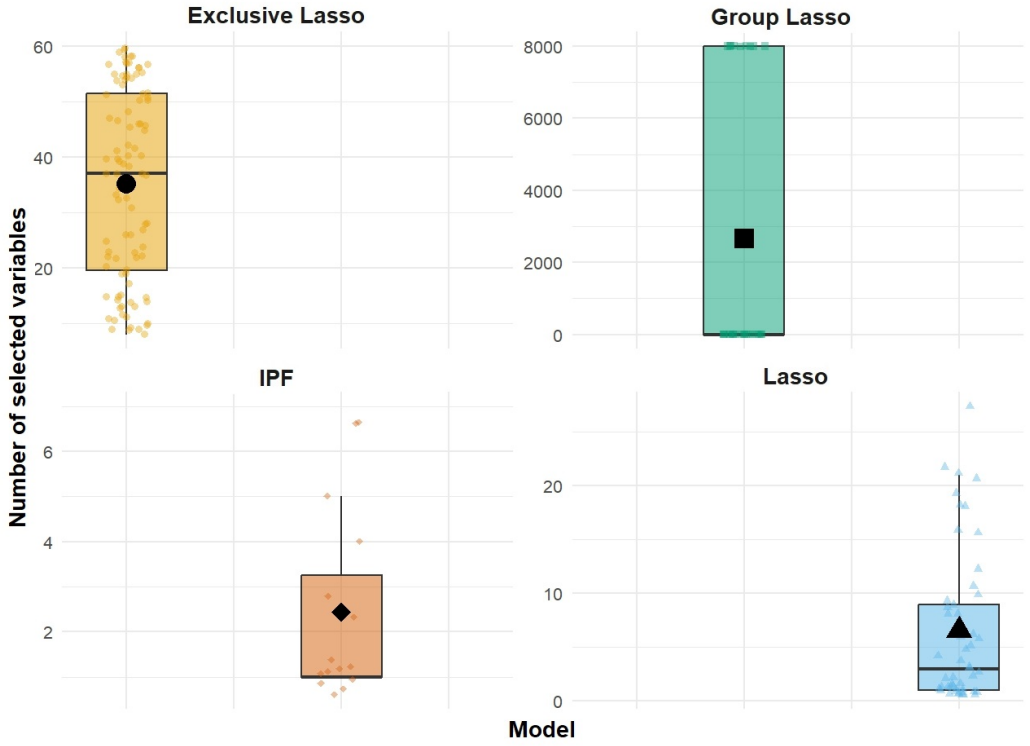

Figure 5: Average number of selected variables across 100 random subsamples for each model in the Bladder cancer gene expression study. The black pointers are the average numbers of selected variables.

consistently selects at least one variable from this group, while other models tend to ignore these variables completely. These variables are only selected by Group Lasso in addition to Exclusive Lasso. In some iterations, Group Lasso selects both clinical and gene expression variables, but this occurs less than 25% of the time (see Figure 5). More frequently, Group Lasso selects only variables from the gene expression group. Group Lasso also has many unique variables that are not selected by any other models, as it selects all variables from a group when that group is chosen. IPF-Lasso selects variables that are shared by all other models, with no unique variables appearing in the top 10 most frequently selected. However, no biomarker is selected more than 10 times out of 100 subsamples. Lasso also fails to select any unique or clinical features, with 5 of the top 10 features being shared by other models.

5 Conclusion

Variable selection plays a critical role in high-dimensional biological datasets. Time-to-event prediction improves when redundant and non-informative features are filtered out, leading to better runtime efficiency and interpretability. However, most filter and prediction methods

fail to account for the intricate grouping structure of biological data. Studies suggest that predictive performance improves when clinical variables are prioritized (Herrmann et al., 2021). However, due to their low dimensionality, clinical variables are often overshadowed by the vast number of gene expression features, particularly when using standard Lasso regularization. To ensure proper representation of low-dimensional clinical variables, we propose using Exclusive Lasso in Cox PH regression models.

The Exclusive Lasso penalty combines the L_1-norm within groups to enforce sparsity among highly correlated features and the L_2-norm between groups to ensure all groups are represented. This approach prevents low-dimensional groups from being overlooked while selecting the most relevant variables within each group, even when they are highly correlated. In contrast, methods like IPF-Lasso account for the grouping structure by applying an L_1-norm within each group but do not guarantee the selection of low-dimensional groups. Additionally, a major drawback of IPF-Lasso is the need to specify a set of penalty factors or weights for each group, which, although potentially data-driven, is a time-consuming process.

In our simulation study, we compared the proposed Exclusive Lasso with other state-of-the-art methods that accounted for grouping structures, such as Elastic Net, Group Lasso, and IPF-Lasso. Exclusive Lasso outperformed these models in terms of selection accuracy and false discovery rate. Although its performance slightly deteriorated as the number of true variables increased, it still maintained a lower false discovery rate than the other models. It also performed well when variables were evenly distributed across groups. While IPF-Lasso achieved comparable performance, it either failed to select variables from certain groups or tended to select highly correlated variables within the same group. Group Lasso, on the other hand, performed poorly as it failed to select variables across all groups.

We analyzed the performance of the methods in a real-world Bladder cancer study. Although the methods had comparable integrated Brier scores, we observed that Exclusive Lasso achieved the best mean Brier score up to 60 months at every time interval. This may be because most methods tend to ignore clinical variables, whereas Exclusive Lasso selects them. The survival prediction and disease progression of bladder cancer are highly influenced by clinical predictors such as tumor stage and smoking status. Therefore, beyond gene selection, incorporating clinical variables into prediction models is crucial. We also found that variable selection in Exclusive Lasso was more consistent across repetitions, whereas other models selected different variables in different iterations.

Although Exclusive Lasso is highly effective in selecting variables from each group, its estimation is challenging due to the composite nature of the penalty. For future work, we are currently developing our proposed *NM*-$L_{1,2}$ algorithm (Ravi and Groll, 2025) for Cox PH models. This method can robustly handle cases where certain groups contain no true variables. Traditional Exclusive Lasso cannot exclude a group even if it is non-informative due to the L_2-norm between groups. To address this, we proposed an ad-hoc technique to include only informative variables. Specifically, we converted the composite norm into a differentiable norm and used a Newton-based algorithm for estimation. We showed that this approach outperforms the coordinate descent method. However, this method needs further

investigation, particularly in scenarios with random allocation. Additionally, we plan to implement stability selection techniques, as proposed by Meinshausen and Bühlmann (2010), to enhance variable selection robustness, especially in cases involving random allocation. We also look forward to testing the method with more levels of grouping, such as multi-omics datasets, where, in addition to clinical variables, multiple layers of omics data—such as genomics, epigenomics, transcriptomics, proteomics, metabolomics, and microbiomics—are available.

Funding This work has been supported by the Research Training Group "Biostatistical Methods for High-Dimensional Data in Toxicology" (RTG 2624, Project R2) funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation—Project Number 427806116).
Competing interests The authors have no competing interests to declare that are relevant to the content of this article.

References

Bommert, A., Welchowski, T., Schmid, M., and Rahnenführer, J. (2022). Benchmark of filter methods for feature selection in high-dimensional gene expression survival data. *Briefings in Bioinformatics*, 23(1):354.

Boulesteix, A.-L., De Bin, R., Jiang, X., and Fuchs, M. (2017). IPF-LASSO: Integrative l1-penalized regression with penalty factors for prediction based on multi-omics data. *Computational and mathematical methods in medicine*, 2017(1):7691937.

Boulesteix, A.-L., Fuchs, M., and Schulze, G. (2019). *IPF-LASSO: Integrative Lasso with Penalty Factors*. R package version 1.1.

Campbell, F. and Allen, G. I. (2017). Within group variable selection through the exclusive lasso. *Electronic Journal of Statistics*, 11:4220–4257.

Cox (1972). Regression models and life tables. *JR Stat Soc*, 34:248–275.

Cumberbatch, M. G. K., Jubber, I., Black, P. C., Esperto, F., Figueroa, J. D., Kamat, A. M., Kiemeney, L., Lotan, Y., Pang, K., Silverman, D. T., et al. (2018). Epidemiology of bladder cancer: a systematic review and contemporary update of risk factors in 2018. *European urology*, 74(6):784–795.

Friedman, J., Tibshirani, R., and Hastie, T. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22.

Graf, E., Schmoor, C., Sauerbrei, W., and Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17-18):2529–2545.

Gregoratti, D., Mestre, X., and Buelga, C. (2021). Exclusive group lasso for structured variable selection. *arXiv preprint arXiv:2108.10284*.

Herrmann, M., Probst, P., Hornung, R., Jurinovic, V., and Boulesteix, A.-L. (2021). Large-scale benchmark study of survival prediction methods using multi-omics data. *Briefings in Bioinformatics*, 22(3):167.

Huang, Y. and Liu, J. (2018). Exclusive sparsity norm minimization with random groups via cone projection. *IEEE Transactions on Neural Networks and Learning Systems*, 29(12):6145–6153.

Kim, J., Sohn, I., Jung, S.-H., Kim, S., and Park, C. (2012). Analysis of survival data with group lasso. *Communications in Statistics-Simulation and Computation*, 41(9):1593–1605.

Kong, D., Fujimaki, R., Liu, J., Nie, F., and Ding, C. (2014). Exclusive feature learning on arbitrary structures via $\ell_{1,2}$-norm. *Advances in Neural Information Processing Systems*, 27.

Lin, M., Yuan, Y., Sun, D., and Toh, K.-C. (2020). Adaptive sieving with ppdna: Generating solution paths of exclusive lasso models. *arXiv preprint arXiv:2009.08719*.

Ma, S., Song, X., and Huang, J. (2007). Supervised group lasso with applications to microarray data analysis. *BMC bioinformatics*, 8:1–17.

Meinshausen, N. and Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 72(4):417–473.

Ravi, D. and Groll, A. (2024). Optimizing variable selection in multi-omics datasets: a focus on exclusive lasso. In *International Workshop on Statistical Modelling*, pages 142–147. Springer.

Ravi, D. and Groll, A. (2025). A newton-based variant of exclusive lasso for improved sparse solutions. *Computational Statistics*. to appear.

Riester, M., Taylor, J. M., Feifer, A., Koppie, T., Rosenberg, J. E., Downey, R. J., Bochner, B. H., and Michor, F. (2012). Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. *Clinical cancer research*, 18(5):1323–1333.

Schulze, G. (2017). *Clinical Outcome Prediction Based on Multi-Omics Data*. PhD thesis.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics*, 22(2):231–245.

Simon, N., Friedman, J., Hastie, T., and Tibshirani, R. (2019). *SGL: Fit a GLM (or Cox Model) with a Combination of Lasso and Group Lasso Regularization*. R package version 1.3.

Simon, N., Friedman, J., Tibshirani, R., and Hastie, T. (2011a). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39(5):1–13.

Simon, N., Friedman, J. H., Hastie, T., and Tibshirani, R. (2011b). Regularization paths for Cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, 39:1–13.

Sun, Y., Chain, B., Kaski, S., and Shawe-Taylor, J. (2020). Correlated feature selection with extended exclusive group lasso. *arXiv preprint arXiv:2002.12460*.

Tibshirani, R. (1997). The lasso method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.

Van Rijsbergen, C. (1979). Information retrieval: theory and practice. In *Proceedings of the joint IBM/University of Newcastle upon tyne seminar on data base systems*, volume 79, pages 1–14.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 68(1):49–67.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine learning research*, 7(Nov):2541–2563.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.