Extending the Distributional Regression Framework:

Treatment Effects, Mixed Responses and Data-driven Variable Selection

---

Dissertation

zur Erlangung des akademischen Grades
doctor rerum naturalium

(Doktor der Naturwissenschaft)

eingereicht an der

Fakultät Statistik

der Technische Universität Dortmund

von

Guillermo Briseño Sanchez, M.Sc.
aus Mexiko

Begutachtet von:
Prof. Dr. Andreas Groll
Prof. Dr. Nadja Klein

Eingereicht: Januar 2025

Abstract

This thesis develops distributional regression methods tailored to the estimation of treatment effects as well as joint modelling of multivariate non-commensurable responses, all based on the Generalised Additive Models for Location Scale and Shape (GAMLSS) approach. In addition, it postulates methods for data-driven variable selection for the aforementioned model class. These developments are introduced across four contributed articles and are implemented in the statistical programming software R.

In the first article, we derive treatment effects on the entire conditional response distribution via an instrumental variable estimation approach based on GAMLSS. Our approach allows to model all parameters of possibly complex outcome distributions as well as non-linear relationships between explanatory variables, instrument and outcome of interest. This demonstrates the potential of using distributional regression in instrumental variable regression both to account for endogeneity and estimate treatment effects beyond the mean.

The second article introduces flexible copula-based statistical models for bivariate responses comprised of non-commensurable (i.e. mixed) variables whose components are a right-censored time-to-event response and a non-time-to-event outcome. The copula approach allows for separate specification of the dependence structure between the margins and their individual distribution functions. The model of the time-to-event margin is constructed via discrete-time-to-event or piecewise-exponential methods using the correspondence of their likelihood of the aforementioned approaches with well-known univariate distributions.

The last two articles tackle the issue of data-driven variable selection for copula-based distributional regression models. In the third article we devise a gradient boosting estimation algorithm adapted to accommodate copula models with arbitrary marginal distributions suited for bivariate binary, count and non-commensurable mixed outcomes. The last article further extends these methods to bivariate right-censored time-to-event responses. This dramatically streamlines the model-building process for a wide range of response structures.

The versatility of the proposed methods is demonstrated through the analysis of various synthetic and real data structures from labour economics, transportation, genetic epidemiology, healthcare utilisation, childhood undernutrition and ovarian cancer.

Zusammenfassung

Diese Dissertation beschäftigt sich mit der Entwicklung von verteilungsregressionsbasierten Methoden, welche zur Schätzung von Treatment-Effekten sowie auf die gemeinsame Modellierung multivariater gemischter Zielgrößen geeignet sind. Alle entwickelten Methoden basieren auf dem Ansatz der Generalised Additive Models for Location Scale and Shape (GAMLSS). Darüber hinaus werden Verfahren zur Variablenselektion für diese Modellklasse entwickelt. Die neuen Verfahren werden in vier wissenschaftlichen Artikeln vorgestellt und in der statistischen Programmiersprache R implementiert.

Im ersten Artikel werden Treatment-Effekte für die gesamte bedingte Verteilung hergeleitet. Dies erfolgt unter Verwendung eines Instrumentalvariablen-Schätzansatzes basierend auf GAMLSS. Unser Ansatz ermöglicht es, alle Parameter möglicher komplexer Verteilungen sowie nichtlineare Beziehungen zwischen erklärenden Variablen, Instrumentalvariablen und die Zielgröße zu modellieren. Dies zeigt die Vorteile, Verteilungsregression in der Instrumentalvariablenregression zu nutzen, um Endogenität zu berücksichtigen und Treatment-Effekte über den Mittelwert hinaus zu schätzen.

Der zweite Artikel führt flexible copula-basierte statistische Modelle für bivariate Zielgrößen ein, die aus gemischten Rändern bestehen. Die Komponenten dieser Zielgrößen bestehen aus einer rechtszensierten Ereigniszeit und einer nicht-Ereigniszeit Variable. Der Copula-Ansatz ermöglicht eine separate Spezifikation der Abhängigkeitsstruktur zwischen den Rändern und ihren individuellen Verteilungsfunktionen. Das Modell der Ereigniszeitkomponente wird mittels discrete time oder piecewise-exponential Verfahren unter Verwendung der Übereinstimmung ihrer Likelihood mit den bekannten univariaten Verteilungen der genannten Ansätze gebildet.

Die letzten zwei Artikel beschäftigen sich mit der Variablenselektion für Copula-basierte Verteilungsregressionsmodelle. Im dritten Artikel wird ein Gradient Boosting-Schätzungsalgorithmus entwickelt. Dieser ist für Copula-Modelle mit beliebigen marginalen Verteilungen angepasst und für bivariate binäre, zähl und gemischte Zielgrößen geeignet. Der vierte Artikel erweitert diese Verfahren auf bivariate rechtszensierte Ereigniszeiten.

Die Vielfältigkeit der Methoden wird durch die Analyse von synthetischen und realen Daten aus verschiedenen Anwendungsgebieten demonstriert.

Acknowledgements

I would like to express my deepest gratitude to Andreas Groll and Nadja Klein for giving me the opportunity to pursue a PhD and develop my research skills. I am very thankful to both for their supervision, guidance and support during my time as a doctoral student.

I want to also extend my gratitude to the other fellow PhD students I met during this time for all the nice moments we had in the office, all the fun we had at the conferences we attended, as well as the memories we made during the multiple retreats that we had. It was a pleasure working with you.

Lastly, and most importantly, I want to sincerely thank my family and my loved ones for their continuous and unconditional love and encouragement during this time, and throughout my life. I am indebted to you for always bringing me strength, joy, and resilience in my life.

Contents

List of figures

List of tables

1 Introduction

Empirical phenomena such as rural unemployment, malnutrition, vehicle choice, poverty, environmental risks, etc. are rarely the consequence of one factor. Instead, they are intrinsically determined by multiple, possibly non-commensurate variables that may exhibit complex dependencies between each other as well as intricate associations with other appended explanatory variables or covariates. The increasing prevalence of multivariate datasets containing these variables and an often overwhelmingly large amount of covariates poses a challenge for conventional statistical analysis tools that are often limited with respect to their flexibility and can only provide a narrow view of the analysed phenomenon. Hence, there is a necessity to devise sophisticated statistical models that accommodate the data's characteristics in order to provide a more accurate picture of the intricate dynamics behind the data.

Regression models are concerned with relating aspects of an outcome variable to covariates. This model class is a prominent tool for empirical analyses. Unfortunately, out-of-the-box or conventional regression techniques are unlikely to be able to capture the complexity of the data. These classical models are bound to simple distributions, strictly linear associations between response and explanatory variables, as well as independence of the outcome variables. Such properties are seldom satisfied or met in practice. In contrast to classical regression, which only considers the response's expectation, *distributional regression* offers a richer framework for empirical analysis by modelling the entire conditional response distribution. In this vein, the investigation of regression effects on the entire distribution may offer a more comprehensive understanding of the variables being analysed.

Motivated by the challenges posed by these complex data architectures, this thesis presents advancements in distributional regression that aim to accommodate the aforementioned response and covariate structures. These methods are tailored to the estimation of treatment effects under endogeneity due to unmeasured confounders, joint modelling of multivariate non-commensurate responses as well as data-driven variable selection. The versatility of the proposed distributional regression methodology is demonstrated through the analysis of various synthetic and real datasets. Additionally, this thesis provides software tools for data analysis. These developments are introduced across four contributed articles.

Recent growing interest to answer questions of causal nature using data from observational studies requires models that not only account for observed as well as unobserved confounders, but also possible dependencies and non-linearities in the data generating process in order to draw accurate causal conclusions. The use of traditional regression techniques in this domain only provides an incomplete or limited understanding of the causal relationship because they are only concerned with changes in the expected outcomes and not their distributions. Consequently, adopting a distributional approach would enrich the conclusions drawn from a causal analysis compared to using classical procedures.

Advances in statistical modelling allow to construct models of phenomena that are determined by multi-dimensional outcomes typically comprised of non-commensurable or mixed components. These models provide an overarching view of the underlying dynamics that determine these empirical processes. A distributional regression approach tailored towards these challenging types of responses would further enhance the depth of the analysis by accounting for possible dependence between the mixed components of the multivariate response.

Lastly, variable selection is a crucial undertaking of regression models. Contemporary data collection methods have made possible the recording and storage of increasingly large datasets. In certain edge cases, the number of explanatory variables related to an observational unit might even exceed the total number of observations, rendering estimation using standard techniques infeasible. Hence, filtering or selecting the most informative variables in order to obtain a sparse and interpretable distributional model of the response of interest should be carefully taken into consideration.

The contributed articles that constitute this dissertation have resulted in the following peer-reviewed publications and unpublished manuscripts that have been submitted at scientific journals and are currently in the review process:

1. **Briseño Sanchez, G.**, Hohberg, M., Groll, A., and Kneib, T. (2020). Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183 (4), 1553–1574.

2. **Briseño Sanchez, G.**, and Groll, A. (2024) Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses. [Manuscript submitted for publication].

3. **Briseño Sanchez, G.**, Klein, N., Klinkhammer, H., and Mayr, A. (2025). Boosting distributional copula regression for bivariate binary, discrete and mixed responses. *Statistical Methods in Medical Research*, https://doi.org/10.1177/09622802241313294.

4. **Briseño Sanchez, G.**, Klein, N., Groll, A., and Mayr, A. (2024). Boosting distributional copula regression for bivariate right-censored time-to-event data. [Manuscript submitted for publication].

In addition, the following peer-reviewed proceedings articles were published during the time of this dissertation:

5. **Briseño Sanchez, G.**, and Groll, A. (2020). Modelling the effect of rural electrification on employment via component-wise boosted causal distributional regression. In I. I. Garbizu, D.-J. Lee, J. M. Minaya, & M. X. R. Álvarez (Eds.) *Proceedings of the 35th International Workshop Statistical Modelling July 20-24, 2020 Bilbao, Basque Country, Spain*, (pp. 25–30). Universidad del País Vasco (Euskal Herriko Unibertsitatea), Argitalpen Zerbitzua.

6. **Briseño Sanchez, G.**, and Groll, A. (2022). Bivariate mixed binary-survival additive regression modelling. In N. Torelli, R. Bellio, and V. Muggeo (Eds.) *Proceedings of the 36th International Workshop Statistical Modelling July 18-22, 2022 - Trieste, Italy*, (pp. 97–102). EUT Edizioni Università di Trieste.

Lastly, I contributed to the following peer-reviewed publication during the time of this dissertation:

7. Marmolejo-Ramos, F., Tejo, M., Brabec, M., Kuzilek, J., Joksimovic, S., Kovanovic, V., González, J., Kneib, T., Bühlmann, P., Kook, L., **Briseño Sanchez, G.**, and Ospina, R. (2023). Distributional regression modeling via generalized additive models for location, scale, and shape: An overview through a data set from learning analytics. *WIREs Data Mining and Knowledge Discovery*, 13 (1), e1479.

All of the methodological developments presented in this thesis have been implemented in the statistical programming software R (R Core Team, 2024). The distributional regression method for causal inference presented in Appendix A can be fitted using any software for distributional regression. A list of R packages and other software is provided at the end of Section 2.1, with a small example

being provided in Appendix E1. Models for mixed responses presented in Appendix B are implemented in a modified version of the R package `GJRM` (Marra and Radice, 2023). The boosting algorithms for estimation and data-driven variable selection introduced in Appendices C and D are implemented in the R package `gamboostLSS` (Hofner et al., 2016). We refer to Appendices E4 and E5 for examples on how to use the implemented routines.

The remainder of this dissertation is structured as follows: Section 2.1 provides an overview of regression methodology and defines the distributional regression framework used throughout the dissertation. Section 2.2 outlines the methods developed for causal inference and estimating distributional treatment effects under endogeneity due to unmeasured confounders. More details may be found in the publication corresponding to this topic in Appendix A. In Section 2.3, we describe how to construct flexible models for non-commensurable outcomes that are comprised of a non-time-to-event margin and a right-censored time-to-event variable. The contributed article corresponding to this topic is found in Appendix B. Data-driven variable selection for distributional regression is discussed in Section 2.4. We refer to the contributed articles in Appendices C and D for detailed descriptions of the proposed methods for variable selection. Section 3 contains a short summary of each of the four contributed articles that constitute this dissertation. Section 4 provides concluding remarks. Lastly, a discussion of potential avenues for future research is given in Section 5.

2 Methodology

In this section we introduce distributional regression for generic univariate and multivariate responses. We refer to each contributed article for more details on the specific type of outcome variables. Appendix A analyses univariate binary and continuous variables. Multivariate responses are considered in Appendices B, C and D. Appendix B is dedicated to non-commensurable outcomes composed of non-time-to-event and time-to-event margins. Bivariate binary, discrete and mixed binary & continuous responses are analysed in Appendix C, whereas Appendix D is concerned with bivariate right-censored time-to-event variables.

2.1 Distributional regression

Regression models consist of a distributional and a structural assumption. The distributional assumption designates a statistical probability distribution for the response, whereas the structural assumption involves the specification of a linear predictor that defines how a response variable depends on a set of covariates. Let Y_i denote the response variable of the i-th observation in a sample of size $n \in \mathbb{N}_+$, with $i = 1, \ldots, n$. In classical linear regression, the distributional assumption states that y_i is a random draw from a homoscedastic Gaussian distribution $\mathcal{N}(\eta_i, \sigma^2)$, where σ^2 denotes the constant variance of the response. The structural assumption relates the conditional expectation to the covariates of interest by a linear relationship. Both assumptions can be summarised as:

$$Y_i|\mathbf{X}_i = \mathbf{x}_i \sim \mathcal{N}(\eta_i, \sigma^2), \qquad \eta_i = \beta_0 + \sum_{r=1}^{P} \beta_r x_{ri}, \tag{1}$$

where β_0 denotes the intercept, the coefficients β_r are linear effects of the covariates x_r in the data, with $r = 1, \ldots, P$. Equation (1) shows that the linear predictor η_i is a model for the conditional expectation of Y_i, i.e. $\mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \eta_i$ (Weisberg, 2014; Fahrmeir et al., 2021). Generalised Linear Models (GLM; Nelder and Wedderburn, 1972), Generalised Additive Models (GAM; Hastie and Tibshirani, 1986; Wood, 2017) and Structured Additive Regression models (STAR; Brezger and Lang, 2006) further extended the range of response types as well as the variety of covariate effects one can specify in the predictor η_i. For instance, GLMs model the conditional expectation of response variables from parametric distributions that belong to the univariate exponential family (EF), encompassing models for binary, count, as well as strictly positive continuous

responses:

$$Y_i|\mathbf{X}_i = \mathbf{x}_i \sim \mathcal{D}_{EF}(\mu_i, \phi), \quad \mathbb{E}(Y_i|\mathbf{X}_i = \mathbf{x}_i) = \mu_i = h(\eta_i) \Leftrightarrow \eta_i = g(\mu_i),$$

where $h(\cdot)$ is a one-to-one, twice differentiable response function, its inverse $g(\cdot)$ is known as link function and ϕ is a dispersion parameter (Fahrmeir et al., 2021). The GAM and STAR frameworks introduced more flexible functional forms of the covariates by replacing the linear predictor with a structured additive predictor of the form:

$$\eta_i = \beta_0 + \sum_{r=1}^{P} s_r(x_{ri}),$$

where $s_r(x_{ri})$ are smooth functions of the covariates, these include linear effects, non-linear functions, spatial effects and random effects (Wood, 2017). In these cases, the response distribution is still restricted to the univariate exponential family. The aforementioned regression techniques only model conditional expectation of a response variable and despite of their flexibility, they describe only one aspect of the outcome's statistical behaviour. Distributional regression addresses this gap by providing a richer understanding of the conditional response distribution.

Generally understood as an umbrella term, distributional regression encompasses a wide range of techniques that model the entire conditional response distribution as a function of covariates (Kneib et al., 2023; Klein, 2024). Well-known examples of distributional regression are quantile regression (Koenker, 2005), expectile regression (Newey and Powell, 1987), Generalised Additive Models for Location Scale and Shape (Rigby and Stasinopoulos, 2005; GAMLSS) and conditional transformation models (Hothorn et al., 2013). The aforementioned techniques provide, to different extents, a complete picture of the conditional response distribution. In particular, GAMLSS relate all of the parameters of the outcome distribution to covariates. This has the advantage that any parametric distribution may be considered as a potential candidate for the outcome distribution, see e.g. Rigby et al. (2019) for a large catalogue of parametric distributions used for GAMLSS modelling. Another important benefit is that the likelihood and other quantities required for estimation (e.g. partial derivatives) from such a model are readily available or can be derived analytically. For these reasons, we focus on distributional regression based on GAMLSS throughout this thesis. Within this framework, it is then possible to construct a very flexible statistical model

for a wide range of response types. Primarily, we focus on univariate and multivariate binary, discrete, and right-censored time-to-event variables, as well as non-commensurable or mixed responses.

Model representation

The key assumption of distributional regression is that the observed responses y_i are conditionally independent random draws from a parametric distribution, we may write this as:

$$Y_i|\mathbf{X}_i = \mathbf{x}_i \sim \mathcal{D}(\boldsymbol{\vartheta}_i),$$

where the K-dimensional vector $\boldsymbol{\vartheta}_i = (\vartheta_{1i}, \ldots, \vartheta_{Ki})^\top$ emphasises the parametric nature of the model and contains all of the parameters that specify the response distribution $\mathcal{D}$. In case of univariate right-censored time-to-event variables, one observes $Y_i = \min\{T_i, T_i^{cens}\}$ accompanied by the censoring indicator $\delta_i = \mathbb{1}\{T_i \leq T_i^{cens}\}$. Here T_i denotes the true event time, T_i^{cens} is a random, non-informative censoring time independent of T_i and $\mathbb{1}\{\cdot\}$ is the indicator function.

When working with multivariate responses, classical parametric multivariate distributions such as the bivariate Gaussian, bivariate Bernoulli or bivariate Poisson distributions may be considered. See Kocherlakota and Kocherlakota (2017) and Lai and Balakrishnan (2009) for a review on bivariate discrete and bivariate continuous distributions, respectively. Major drawbacks of classical multivariate distributions are the implicit restrictions made on the type of marginal distributions and dependence structure. For example, the bivariate Gaussian distribution imposes Gaussian margins, whereas the bivariate Poisson distribution imposes Poisson margins and is only able to model positive dependence between them. We refer to Klein et al. (2014), Groll et al. (2018), and Strömer et al. (2023) for more details on multivariate distributional regression based on classical multivariate distributions. Classical multivariate distributions are too restrictive whenever the marginal responses exhibit complex dependence structures or happen to follow different families of distributions, for example one margin is Gaussian and the other is Weibull distributed. If the bivariate or multivariate response of interest consists of non-commensurable margins, i.e. margins of different or mixed type, for example an outcome determined by binary and continuous margins, then classical parametric multivariate distributions are rendered unsuitable for modelling. In such cases, the joint distribution may be instead constructed by means of a copula function (Nelsen, 2006).

Let $\boldsymbol{Y}_i = (Y_{1i}, Y_{2i})$ be the i-th bivariate response with corresponding parametric marginal cumulative distribution functions (CDF) $F_1(y_{1i};\ \boldsymbol{\vartheta}_i^{(1)})$, and $F_2(y_{2i};\ \boldsymbol{\vartheta}_i^{(2)})$. Using the copula approach, the joint distribution of $\boldsymbol{Y}_i$ denoted by $P(Y_{1i} \leq y_{1i}, Y_{2i} \leq y_{2i}; \boldsymbol{\vartheta}_i) = F(y_{1i},\ y_{2i};\ \boldsymbol{\vartheta}_i)$, may be written as:

$$F(y_{1i},\ y_{2i};\ \boldsymbol{\vartheta}_i) = C[F_1(y_{1i};\ \boldsymbol{\vartheta}_i^{(1)}),\ F_2(y_{2i};\ \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}], \tag{2}$$

where $C(\cdot, \cdot) : [0, 1]^2 \to [0, 1]$ is the CDF of a parametric bivariate copula with dependence parameter $\vartheta_i^{(c)}$ that binds the margins together. In a similar fashion, the joint survival function of a bivariate time-to-event outcome denoted by $P(T_{1i} > t_{1i}, T_{2i} > t_{2i}; \boldsymbol{\vartheta}_i) = S(t_{1i},\ t_{2i};\ \boldsymbol{\vartheta}_i)$ can be constructed using a copula function and the marginal survival functions $S_1(t_{1i}; \boldsymbol{\vartheta}_i^{(1)})$ and $S_2(t_{2i}; \boldsymbol{\vartheta}_i^{(2)})$:

$$S(t_{1i},\ t_{2i};\ \boldsymbol{\vartheta}_i) = C[S_1(t_{1i};\ \boldsymbol{\vartheta}_i^{(1)}),\ S_2(t_{2i};\ \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}]. \tag{3}$$

Equations (2) and (3) once again emphasise the parametric nature of the bivariate distribution and the joint survival function, respectively. The copula allows to decompose the statistical model into arbitrary, albeit suitable marginal distributions and the dependence structure between them, which is completely described by the scalar parameter $\vartheta_i^{(c)} \in \Theta^{(c)}$. The range of the dependence parameter $\Theta^{(c)}$ depends on the assumed parametric copula family. However, values of $\vartheta_i^{(c)}$ can be transformed to Kendall's $\tau \in [-1; +1]$ or Spearman's $\rho \in [-1; +1]$ in order to better interpret the strength of the dependence between the margins. Throughout this thesis we refer to models for multivariate responses based on Equations (2) or (3) as "copula-based distributional regression" or "distributional copula regression" interchangeably.

In the context of bivariate responses, the $K = K_1 + K_2 + 1$ dimensional parameter vector $\boldsymbol{\vartheta}_i$ encompasses the parameter vectors that correspond to the marginal distributions $\boldsymbol{\vartheta}_i^{(1)} \in \mathbb{R}^{K_1}, \boldsymbol{\vartheta}_i^{(2)} \in \mathbb{R}^{K_2}$, and the scalar $\vartheta_i^{(c)}$, i.e. $\boldsymbol{\vartheta}_i = ((\boldsymbol{\vartheta}_i^{(1)})^\top, (\boldsymbol{\vartheta}_i^{(2)})^\top, \vartheta^{(c)})^\top$. Throughout this thesis one-parameter copulas are considered. However, more complex copulas with a vector-valued dependence parameter such as the Student's t copula with $\boldsymbol{\vartheta}^{(c)} = \left(\vartheta_1^{(c)}, \vartheta_2^{(c)}\right)$ can be easily embedded into this framework, see for example Marra and Radice (2017) and Marra and Radice (2020).

Sklar's Theorem guarantees that the copula is uniquely defined when continuous margins are used (Nelsen, 2006). When discrete margins are present, the copula

is uniquely defined only on the range of the marginal CDFs. However, within our framework, identifiability should not be an issue for two reasons. First, the parametric form of the joint distribution is determined *a priori* by making choices for the marginal distributions and the copula function. Hence, potential identifiability issues that arise when one is interested in estimating e.g. the copula and the marginals in a nonparametric framework without an *a priori* fixed structure, are not present. Second, identifiability is ensured in our regression setting where all parameters of the distribution are observation-specific. Consider two observations, say i and i', with the same observed marginal response ($y_{ji} = y_{ji'}$) but different covariate values for $j = 1, 2$. Modelling the parameters of the respective marginal distributions as functions of covariates results in different estimates for the marginal CDFs, i.e. $\hat{F}_j(y_{ji}; \hat{\boldsymbol{\vartheta}}_i) \neq \hat{F}_j(y_{ji'}; \hat{\boldsymbol{\vartheta}}_{i'}), j = 1, 2$. Therefore, a richer range of the estimated CDFs of the discrete marginal distributions is obtained, mitigating the identification issue when using these type of marginal responses. This has also been pointed out in other instances in the literature, see e.g. Nikoloulopoulos and Karlis (2010); Joe (2014); Marra and Wyszynski (2016) and Wyszynski and Marra (2018).

Structured additive predictors

Each of the K parameters of the univariate or multivariate response distribution, i.e. entries of the vector $\boldsymbol{\vartheta}_i$, is modelled as a function of covariates $\mathbf{x}_i \in \mathbb{R}^P$ using a structured additive predictor $\eta_{ik}^{(\bullet)}$:

$$\eta_{ik}^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} s_{rk}^{(\bullet)}(x_{ir}), \qquad \text{with } \bullet \in \{1, 2, c\}, \tag{4}$$

where the symbol $(\bullet) \in \{1, 2, c\}$ indicates membership of a quantity to the first margin, the second margin, or the copula dependence parameter in case that multivariate responses are being modelled. In order to guarantee that the individual parameters comply with their respective parameter space restrictions, a suitable response function $h_k^{(\bullet)}(\cdot)$ with corresponding inverse function $g_k^{(\bullet)}(\cdot)$ (known as link), i.e. $h_k^{(\bullet)}(\cdot) \equiv g_k^{-1(\bullet)}(\cdot)$, is applied to each additive predictor $\eta_{ik}^{(\bullet)}$:

$$\vartheta_{ik}^{(\bullet)} = h_k^{(\bullet)}(\eta_{ik}^{(\bullet)}) \in \Theta_k^{(\bullet)} \Longleftrightarrow g_k^{(\bullet)}(\vartheta_{ik}^{(\bullet)}) = \eta_{ik}^{(\bullet)} \in \mathbb{R}, \quad \text{with } \bullet \in \{1, 2, c\},$$

where $\Theta_k^{(\bullet)}$ denotes the corresponding parameter space of $\vartheta_{ik}^{(\bullet)}$. The structured additive predictors $\eta_k^{(\bullet)}$ are composed of a parameter-specific intercept $\beta_{0k}^{(\bullet)}$ and

smooth functions of the covariates denoted by $s_{rk}^{(\bullet)}(\cdot)$. The latter can accommodate a wide range of functional forms, such as linear, non-linear, and spatial effects. This great flexibility is achieved by casting each smooth function $s_{rk}^{(\bullet)}(\cdot)$ through a linear combination of appropriate basis function expansions of the form:

$$s_{rk}^{(\bullet)}(\boldsymbol{x}_{irk}) = \sum_{l=1}^{L_{rk}^{(\bullet)}} \beta_{rk,l}^{(\bullet)} B_{rk,l}^{(\bullet)}(x_{irk}), \quad \text{with } \bullet \in \{1, 2, c\},$$

where $B_{rk,l}^{(\bullet)}(x_{irk})$ are basis functions evaluated at x_{irk}, and $\beta_{rk,l}^{(\bullet)}$ are the corresponding unknown regression coefficients which must be estimated. The above formulation allows one to employ a rich variety of covariate effects, such as linear or non-linear functional forms. See Wood (2017) for more details on the specifications of the smooth functions. Note that for some covariate effects, the unknown coefficients may be associated with a quadratic penalty and a scalar smoothing parameter in order to enforce smoothness of the estimated functions, for example non-linear effects using P-splines (Eilers and Marx, 1996). The penalty term of the r-th smooth covariate effect may be expressed as:

$$Pen(\boldsymbol{\beta}_{rk}^{(\bullet)}) = \lambda_{rk}^{(\bullet)} \boldsymbol{\beta}_{rk}^{(\bullet)^\top} \boldsymbol{D}_{rk}^{(\bullet)} \boldsymbol{\beta}_{rk}^{(\bullet)},$$

where $\boldsymbol{\beta}_{rk}^{(\bullet)} = (\beta_{rk,1}^{(\bullet)}, \ldots, \beta_{rk,L_{rk}^{(\bullet)}}^{(\bullet)})^\top$ is the vector of unknown coefficients associated to the r-th smooth function, $\boldsymbol{D}_{rk}^{(\bullet)}$ is a penalty matrix of suitable dimensions and $\lambda_{rk}^{(\bullet)}$ is a scalar smoothing parameter, with $r = 1, \ldots, P_k^{(\bullet)}$, $k = 1, \ldots, K_{(\bullet)}$, and $\bullet \in \{1, 2, c\}$, see Wood (2017) for more details.

The summation limit $P_k^{(\bullet)}$ in Equation (4) emphasises that the subset of covariates assigned to each parameter do not need to be the same. In fact, it may be the case that no covariates have an effect on some parameters $\vartheta_k^{(\bullet)}$ of the univariate distribution $F(\cdot, \boldsymbol{\vartheta})$, the joint distribution $F(\cdot, \cdot; \boldsymbol{\vartheta})$ or the joint survival function $S(\cdot, \cdot; \boldsymbol{\vartheta})$, depending on the response type being considered. This implies that variable selection is a task of utmost importance for the considered model class. The topic is further discussed in Section 2.4.

Estimation and software

Estimation of the unknown regression coefficients can be conducted using penalised maximum likelihood, Bayesian inference or statistical boosting. For univariate responses, the `R` packages `gamlss` (Stasinopoulos et al., 2024) and `GJRM` (Marra and Radice, 2023) contain estimation routines based on penalised maxi-

mum likelihood. Various distributional regression models have also been implemented in `mgcv` (Wood, 2023). The package `bamlss` (Umlauf et al., 2024) as well as the software BayesX (Belitz et al., 2024) and Liesl (Riebl et al., 2022) rely on fully Bayesian inference. The packages `mboost` (Hothorn et al., 2022) and `gamboostLSS` (Hofner et al., 2016) use statistical boosting.

In terms of multivariate responses, `GJRM`, `gamCopula` (Nagler and Vatter, 2022) and `VGAM` (Yee and Moler, 2024) fit models based on copulas as well as traditional multivariate distributions using penalised maximum likelihood. BayesX allows for copula-based models as well as various multivariate distributions, whereas `bamlss` implements the multivariate Gaussian distribution. The packages `gamboostLSS` and `boostcopula` (Jobst et al., 2024) implement boosting routines for copula-based multivariate distributional regression models.

In this thesis we resort to penalised maximum likelihood as well as statistical boosting to estimate the unknown regression coefficients. The univariate distributional regression models considered in Appendix A are estimated using penalised maximum likelihood using the algorithm implemented in the package `GJRM`. The copula-based distributional regression models for non-commensurable outcomes introduced in Appendix B are fitted via penalised maximum likelihood using the simultaneous estimation algorithm implemented in the `GJRM` package. Lastly, the distributional copula regression models presented in Appendices C and D estimate the unknown model coefficients using statistical boosting based on the implementation of the `gamboostLSS` package.

2.2 Distributional treatment effects

This subsection provides an overview of Two-Stage GAMLSS (2SGAMLSS), an instrumental variable (Stock, 2001; IV) estimation procedure within the distributional regression framework introduced in the contributed article in Appendix A. The purpose of this IV distributional regression approach is to estimate treatment effects on the entire conditional outcome distribution. The proposed method extends the scope of IV regression towards applications dealing with distributional questions that can be answered using *one* statistical model.

In cases where randomisation is not possible, treatment assignment is biased due to self-selection or other sources of endogeneity. Hence, in many observational studies, quasi-experimental settings, experimental settings with low compliance, or when an explanatory or treatment variable is suspected to be endogenous

due to unmeasured confounders, an instrumental variable (x_{IV}) is still able to recover a causal effect. Such endogenous variables are usually denoted by x_{en}. The source of endogeneity is typically linked to omitted variables (unmeasured confounders x_u) that are associated with the treatment as well as the response. Other covariates in the data that are independent of the unmeasured confounders are called exogenous and are denoted by x_{ex}. See Figure 1 for an illustration. The proposed 2SGAMLSS is based on the two-stage residual inclusion approach (Terza et al., 2008), a control function technique (Wooldridge, 2015) that has previously been applied to GAMs (Marra and Radice, 2011). The remainder of this subsection introduces 2SGAMLSS and illustrates the type of treatment effects than can be derived from it. We refer to Appendix A for more details and an application of 2SGAMLSS to labour economics data.

Two-Stage GAMLSS in instrumental variable regression (2SGAMLSS)

Two-stage GAMLSS (2SGAMLSS) consists of two estimation steps. First, a distributional regression model is fitted to the endogenous variable x_{en}:

$$\eta_{ik}^{[1]} = \beta_{0k}^{[1]} + s_{IV}^{[1]}(x_{IV,i}) + \sum_{r=1}^{P_k^{[1]}} s_{rk}^{[1]}(x_{irk}), \qquad k = 1, \ldots, K_{[1]}, \tag{5}$$

where $\eta_{ik}^{[1]}$ is the structured additive predictor of the k-th parameter of x_{en}'s distribution and $i = 1, \ldots, n$. The corresponding parameter is obtained by applying a suitable response function as described in Section 2.1. Here, [1] indicates that the terms specified in Equation (5) belong to the first-stage model. The structured additive predictor contains an intercept, as well as effects for the instrument x_{IV} and the remaining exogenous regressors. Note that in the case of binary endogenous variables ($x_{en} \in \{0, 1\}$), the amount of predictors in the first stage will usually be one, i.e. $K_{[1]} = 1$, since x_{en} will be modelled using a Bernoulli distribution. When complex treatments expressed as continuous or discrete variables are considered, the value of $K_{[1]}$ might be larger than one, depending on the parametric distribution assumed for x_{en}. After fitting the first-stage model, the conditional expectation of the endogenous regressor and the residuals $\hat{\xi}$ are computed:

$$\hat{\xi}_i = x_{en,i} - \mathbb{E}(x_{en,i} \mid \hat{\boldsymbol{\vartheta}}_i^{[1]}), \tag{6}$$

where $\hat{\boldsymbol{\vartheta}}_i^{[1]}$ is the vector of estimated parameters of the distribution of x_{en} obtained

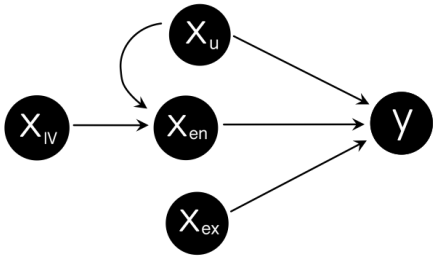

Figure 1: Relationship between response (y), endogenous treatment (x_{en}), unmeasured confounders (x_u), exogenous variables (x_{ex}) and an instrument (x_{IV}).

in the first stage. Subsequently, all $K_{[2]}$ parameters of the response's density $f(y_i;\, \boldsymbol{\vartheta}_i^{[2]})$ are regressed on the explanatory variables and the first-stage residuals using structured additive predictors:

$$\eta_{ik}^{[2]} = \beta_{0k}^{[2]} + s_{en}^{[2]}(x_{en,i}) + s_{\hat{\xi}}^{[2]}(\hat{\xi}_i) + \sum_{r=1}^{P_k^{[2]}} s_{rk}^{[2]}(x_{irk}), \qquad k = 1, \ldots, K_{[2]}. \tag{7}$$

Here, [2] indicates that the terms of Equation (7) belong to the second-stage model and $s_{\hat{\xi}}^{[2]}(\hat{\xi}_i)$ denotes the smooth function of the first-stage residuals. Note that extending this framework to multiple endogenous regressors results in multiple first-stage models, and having all first-stage residuals attached to the structured additive predictors in Equation (7). In the publication contained in Appendix A, 2SGAMLSS was fitted using the routine `gamlss()` included the package `GJRM`. We remark that 2SGAMLSS may be fitted using any routine for distributional regression like those mentioned at the end of Section 2.1. See Section 5 for a discussion on fitting 2SGAMLSS via statistical boosting.

Treatment effects derived from 2SGAMLSS

Causal effects of a binary treatment $D \in \{0, 1\}$ on an outcome $Y(D)$ are defined as differences between the potential outcomes under treatment $Y_1 = Y(1)$ and without treatment $Y_0 = Y(0)$ (Neyman, 1990; Rubin, 1974). However, one is only able to observe either Y_1 or Y_0, depending on the treatment status assigned to the observational unit. The conditional average treatment effect (ATE) reflects the impact of treatment on the expected value of the outcome within a heterogeneous population characterised by covariates:

$$\text{ATE}(\mathbf{x}) = E(Y_1 \mid \mathbf{X} = \mathbf{x}) - E(Y_0 \mid \mathbf{X} = \mathbf{x}).$$

Using 2SGAMLSS, the entire conditional response distribution is estimated using

one model. This allows for the investigation of general distributional aspects of the response such as variance, skewness, etc. For example in income inequality, an intervention that lowers the variance or the Gini coefficient of an income distribution over one that has a similar ATE but does not reduce inequality might be preferred. In such cases the ATE would provide a narrow view on the treatment effect on the outcome. The proposed 2SGAMLSS provides a broader perspective by considering potential changes in the complete conditional response distribution. Therefore, there is not one single scalar treatment effect but rather several treatment effects on various aspects of the response distribution. Suppose that the response Y follows a Gaussian distribution regardless of whether it receives the treatment or not. However, the parameters of the distribution differ by the treatment as well as the covariates, i.e.

$$Y_1|\mathbf{X}=\mathbf{x} \sim \mathcal{N}\left(\vartheta_{1,Y_1}, \vartheta_{2,Y_1}\right) \quad \text{and} \quad Y_0|\mathbf{X}=\mathbf{x} \sim \mathcal{N}\left(\vartheta_{1,Y_0}, \vartheta_{2,Y_0}\right),$$

where due to the Gaussian assumption the distribution parameters ϑ_{1,Y_1} and ϑ_{2,Y_1} denote the expectation and standard deviation of Y with treatment, respectively. The same statement applies for Y without treatment. We can then derive a treatment effect (TE) on any distributional quantity φ by evaluating the difference in φ under different treatment values:

$$\mathrm{TE}_{\varphi}(\mathbf{x}, d_0, d_1) = \varphi(\mathbf{x}, d_1) - \varphi(\mathbf{x}, d_0), \tag{8}$$

where $\mathbf{x}$ are some covariates, d_0 denotes no treatment and d_1 denotes treatment. Assuming a binary treatment variable, the TE on the expected value μ is given by the difference in expected values. The TE on the standard deviation σ can be obtained in a similar fashion. Further TEs on other quantities such as the coefficient of variation σ/μ may be investigated as well:

$$\begin{aligned}
\mathrm{TE}_{\mu}(\mathbf{x}) &= \vartheta_{1,Y_1} - \vartheta_{1,Y_0}, \\
\mathrm{TE}_{\sigma}(\mathbf{x}) &= \vartheta_{2,Y_1} - \vartheta_{2,Y_0}, \\
\mathrm{TE}_{\sigma/\mu}(\mathbf{x}) &= \frac{\vartheta_{2,Y_1}}{\vartheta_{1,Y_1}} - \frac{\vartheta_{2,Y_0}}{\vartheta_{1,Y_0}},
\end{aligned}$$

or on the quantiles of the distribution by evaluating the inverse cumulative distributions functions of the corresponding estimated Gaussian distributions. In cases with complex treatments expressed as continuous or discrete variables, the setup from Equation (8) remains the same, but the status or levels of the treat-

ment variable D before treatment ($D = d_0$) and after treatment ($D = d_1$) have to be determined beforehand. In those cases, the sample mean or other representative values of interest may be considered as original treatment status d_0. Such treatment effects usually explicitly depend on both values d_0 and d_1. For example, when the treatment is given by the dose of a drug, the treatment effect might change depending on the considered dose levels or sizes d_0 and d_1.

The proposed IV distributional regression method provides a variety of treatment effects on various distributional features that can be derived from the estimated parameters of the outcome distribution with and without treatment. The benefit of 2SGAMLSS becomes more evident when the assumed Gaussian response distribution is replaced with more general types of parametric distributions that are determined by multiple parameters. In such cases, expectation, variance, and other distributional features might depend jointly on said parameters. This highlights the necessity and benefit of a distributional regression approach.

2.3 Bivariate mixed non-time-to-event and time-to-event outcomes

This subsection is dedicated to distributional regression models for bivariate non-commensurable (i.e. mixed) outcomes. The margins of the mixed outcomes of interest are comprised of a non-time-to-event variable and a right-censored time-to-event variable, see Figure 2 for three examples. For this subsection and without loss of generality, we write the non-time-to-event variable as Y and the time-to-event margin as T. Our primary interest is to embed a flexible semi-parametric regression model of the time-to-event margin's hazard rate into our copula-based distributional regression framework. We provide an approach to achieve this goal built on functions derived from Bernoulli and Poisson likelihoods. The model for the non-time-to-event margin Y is of secondary interest for the remainder of this subsection. However, we remark that it is specified using structured additive predictors as shown in Equation (4). The proposed copula-based distributional regression models for the aforementioned mixed outcomes have been implemented in a modified version of the R package `GJRM` (Marra and Radice, 2023; version 0.2-5). Appendix E2 provides details on the modifications made to the package as well as demonstrations on how to use the software. Note that the index $i = 1, \ldots, n$ which denotes the observational unit is omitted for the remainder of this subsection in order to avoid clutter in the notation.

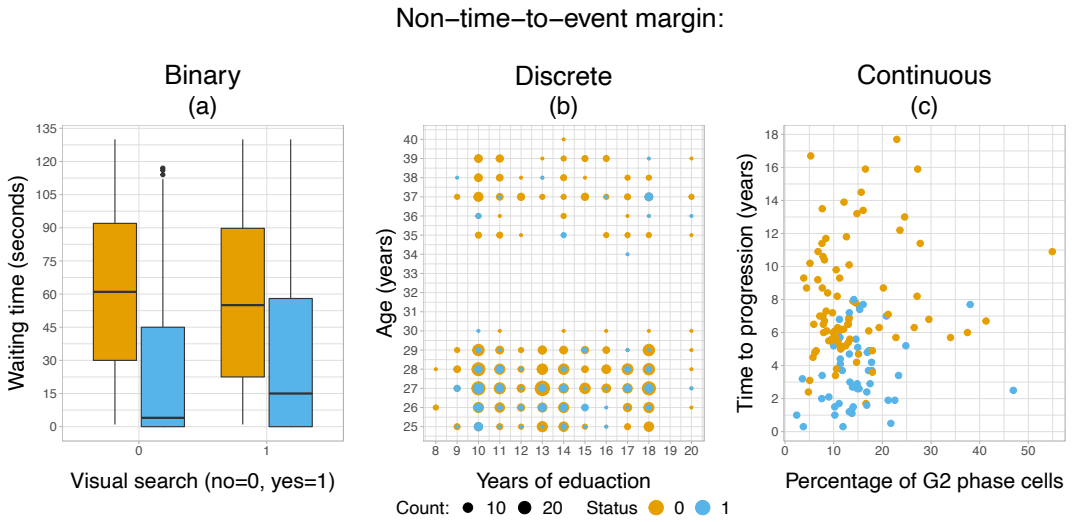

Figure 2: Examples of bivariate non-commensurable outcomes with non-time-to-event & time-to-event margins. Visual search & waiting time (a). Years of education & age at first pregnancy or child birth (b). Percentage of cells in G2 phase & time to tumour progression (c).

Discrete time and piecewise-exponential approaches

In order to construct a flexible model of the hazard rate corresponding to the time-to-event margin, we resort to discrete time-to-event (DT) as well as piecewise-exponential (PW) techniques. Recall that the event and censoring times are denoted by T and T^{cens}, respectively, and in case of right-censored data one observes $\tilde{T} = \min\{T, T^{cens}\}$ with its censoring indicator $\delta = \mathbb{1}\{T \leq T^{cens}\}$. We assume that the censoring time T^{cens} is independent and non-informative of the event time T. Albeit it is possible to relax the assumption of independent censoring and allow for dependence between T and T^{cens}, this will result in a more complex model that the one introduced here. We refer to Czado and Van Keilegom (2022) for more details on models for dependent censoring using copulas.

Let $\mathcal{I}_j$ with $j = 1, \ldots, J$ denote an interval of finite length with respective left and right bounds κ_{j-1}, κ_j, and $\kappa_0 = 0 < \kappa_1 < \cdots < \kappa_J < \infty$. The time axis is then dichotomised into J intervals $\mathcal{I}_j$ with $j = 1, \ldots, J$ using the aforementioned bounds. In the DT framework, the time-to-event variable T takes values in $\{1, 2, \ldots, J\}$, indicating the interval in which an event or censoring occurred. The conditional discrete hazard rate (Tutz and Schmid, 2016) is defined as the

following conditional probability:

$$\lambda(t \mid \mathbf{x}) = P(T = t \mid T \geq t,\ \mathbf{x}) \ \in [0, 1]. \tag{9}$$

In the *PW* approach (Holford, 1980; Laird and Olivier, 1981), the hazard rate is defined as

$$\lambda(t \mid \mathbf{x}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t,\ \mathbf{x}) \ \geq 0, \tag{10}$$

and it is assumed to be constant within each interval. Figure 3 illustrates how the underlying hazard and survival functions are approximated using the piecewise constant functions from the *DT* and *PW* approaches.

Upon closer inspection, the *DT* and *PW* hazard functions from Equations (9) and (10) are only required to obey their respective range restrictions. Therefore, a structured additive predictor can be set as foundation for a flexible regression model for either hazard:

$$\lambda(t \mid \mathbf{x}) = h(\underbrace{s_0(t)}_{\text{link-scale baseline hazard}} + \sum_{r=1}^{P} s_r(t, x_r)) \quad \forall\, t \in\ \mathcal{I}_j,$$

where the baseline hazard is modelled as a smooth function of time via P-splines (Eilers and Marx, 1996) and the functional form of the covariates can be any of those described in Section 2.1. The adopted notation for the smooth effects emphasizes that the covariate effects may also depend on time. The choice of response function $h(\cdot)$ determines whether a *DT* or *PW* hazard is being modelled. The exponential function yields the *PW* hazard, whereas the logistic, inverse-probit and inverse-clog-log functions produce *DT* hazard probabilities. Any other suitable response function for parameters that lie in the unit interval ($\vartheta \in [0, 1]$) may be used in the *DT* case.

In order to achieve correspondence with the Bernoulli and Poisson log-likelihoods, we transform the data into a "long format". Consider $n = 3$ hypothetical subjects with event or censoring times shown in Table 1. For simplicity, we use integer times, but the augmentation applies to continuous times in the same fashion. We assume $J = 4$ intervals with cut points at $\kappa_0 = 0, \kappa_1 = 1.5, \kappa_2 = 3, \kappa_3 = 4$ and $\kappa_4 = 7$ (note that the intervals $\mathcal{I}_j$ can also be equidistant). Subject $i = 1$ was censored in $\mathcal{I}_1$, whereas $i = 2$ experienced an event in $\mathcal{I}_4$, and $i = 3$ was

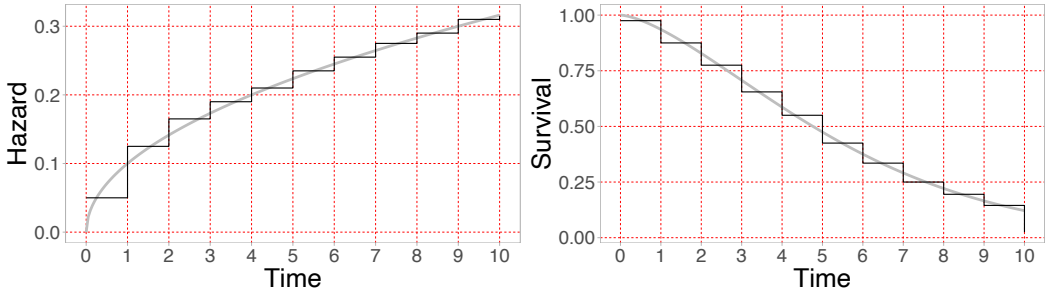

Figure 3: Illustration of the approximation of the hazard and survival functions by means of piecewise-constant functions.

censored in $\mathcal{I}_2$. Column three of the augmented data in Table 1 shows the auxiliary variables δ_{ij} and column four shows the covariates of subject i at interval j, making the inclusion of time-varying covariates straightforward using these techniques. Columns five and six (t_{ij} & o_{ij}) show the time observation i spent in interval j, defined as $t_{ij} = \min\{t_i - \kappa_{j-1}, \kappa_j - \kappa_{j-1}\}$. For example, subject $i = 1$ was censored at $t_1 = 1$, hence $t_{11} = t_1 - \kappa_0 = 1 - \kappa_0 = 1 - 0 = \ 1$. This data augmentation produces the following auxiliary variable:

$$\delta_j = \begin{cases} 1, & \text{if subject has an event in } \mathcal{I}_j, \\ 0, & \text{else,} \end{cases} \quad \text{with} \quad \mathcal{I}_j = \begin{cases} [\kappa_{j-1},\ \kappa_j) & \text{if } DT, \\ (\kappa_{j-1},\ \kappa_j] & \text{if } PW, \end{cases}$$

where κ_{j-1}, κ_j denote the left and right bounds of the intervals used in either DT or PW approaches to dichotomise the time axis. Using the transformed data, the log-likelihood of a model for the DT hazard probability coincides with the Bernoulli log-likelihood (Tutz and Schmid, 2016). The hazard probability at the j-th interval is then $\lambda(j \mid \mathbf{x}_j) = \vartheta_j = g^{-1}(\eta_j)$, where $\vartheta_j \in [0, 1]$ is the parameter of the Bernoulli distribution. In the PW case, the correspondence of the model's log-likelihood is with the Poisson log-likelihood with a given offset o_j, which indicates the time spent in the j-th interval (Bender et al., 2018). In this case, the hazard rate at the j-th interval is $\lambda(j \mid \mathbf{x}_j) = \exp(\eta_j) = \vartheta_j / \exp(o_j)$, where $\vartheta_j = \exp(\eta_j) \exp(o_j)$ is the parameter of the Poisson distribution.

Embedding DT and PW approaches into distributional regression

The i-th time-to-event response is now represented by $j = 1, \ldots, j(i)$ auxiliary variables, where $j(i)$ denotes the length of the sequence of auxiliary variables emanating from the i-th observational unit in the sample. We denote the i-th sequence as $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_{j(i)})^\top$ which can have two possible expressions: $\boldsymbol{\delta} = \mathbf{0}_{j(i)}^\top$

Table 1: Illustration of the data augmentation for *DT* and *PW* approaches using four intervals with borders: $\kappa_0 = 0,\ \kappa_1 = 1.5,\ \kappa_2 = 3,\ \kappa_3 = 4$ and $\kappa_4 = 7$. The column header in the augmented data indicates approach-specific columns.

Original data			
i	t_i	δ_i	$\mathbf{x}_i$
1	1	0	$\mathbf{x}_1$
2	5	1	$\mathbf{x}_2$
3	2	0	$\mathbf{x}_3$

$\Longrightarrow$

Augmented data					
		DT & PW			*PW*
i	j	δ_{ij}	$\mathbf{x}_{ij}$	t_{ij}	$o_{ij} = \ln(t_{ij})$
1	1	0	$\mathbf{x}_{11}$	1	log(1)
2	1	0	$\mathbf{x}_{21}$	1.5	log(1.5)
2	2	0	$\mathbf{x}_{22}$	1.5	log(1.5)
2	3	0	$\mathbf{x}_{23}$	1	log(1)
2	4	1	$\mathbf{x}_{24}$	1	log(1)
3	1	0	$\mathbf{x}_{31}$	1.5	log(1.5)
3	2	0	$\mathbf{x}_{32}$	0.5	log(0.5)

in case of censoring, or $\boldsymbol{\delta} = (\mathbf{0}_{j(i)-1}^\top,\ 1)^\top$ in case of an event, where $\mathbf{0}_{j(i)}^\top$ and $\mathbf{0}_{j(i)-1}^\top$ denote zero vectors of length $j(i)$ and $j(i)-1$, respectively.

We propose to use the following function to model the hazard rate of the time-to-event variable using either *DT* or *PW* techniques:

$$F(\boldsymbol{\delta}) = \begin{cases} f\left(\mathbf{0}_{j(i)}\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^\top \\ f\left(\mathbf{0}_{j(i)}\right) + f\left(\left(\mathbf{0}_{j(i)-1}, 1\right)\right), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1}, 1\right)^\top, \end{cases} \tag{11}$$

where the expression of $f(\boldsymbol{\delta})$ depends on the adopted approach. For instance, opting for a *DT* model and setting the link function to the clog-log function, i.e. $\lambda(j|\mathbf{x}) = 1 - \exp(-\exp(\eta_j))$ results in the following expression of $f(\boldsymbol{\delta})$ based on the Bernoulli likelihood:

$$f(\boldsymbol{\delta}) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^\top, \\ \exp\left(-\sum_{j=1}^{j(i)-1} \exp(\eta_j)\right) - \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1}, 1\right)^\top. \end{cases}$$

If a *PW* model is preferred, the conditional hazard corresponds to $\lambda(j|\mathbf{x}) =$

$\exp(\eta_j)$, which leads to the expression of $f(\boldsymbol{\delta})$ based on the Poisson likelihood:

$$f(\boldsymbol{\delta}) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j + o_j)\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^{\top}, \\ \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j + o_j)\right) \exp(\eta_{j(i)} + o_{j(i)}), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1}, 1\right)^{\top}, \end{cases}$$

where o_j is the offset. We provide analytical first and second order partial derivatives of $f(\boldsymbol{\delta})$ and $F(\boldsymbol{\delta})$ based on DT and PW approaches w.r.t. a generic coefficient vector $\boldsymbol{\beta}$. These derivatives are required for the penalised maximum likelihood simultaneous estimation algorithm of the software implementation.

Model representation and considerations for dependence modelling

Our goal is to construct a model for the joint probability of the non-time-to-event variable Y and the time-to-event margin T (occurring after t). In the context of copula-based distributional regression this may be written as

$$P(Y \leq y,\ T > t;\ \boldsymbol{\vartheta}) = C\left(F_1(y;\ \boldsymbol{\vartheta}^{(1)}),\ S_2(t;\ \boldsymbol{\vartheta}^{(2)})\ ;\ \vartheta^{(c)}\right),$$

where $F_1(\cdot)$ is the CDF of the non-time-to-event margin and $S_2(\cdot)$ is the survival function of the time-to-event variable T. We now replace the survival function of the potentially right-censored time-to-event margin with the proposed function $F(\boldsymbol{\delta})$ from Equation (11). We focus primarily on cases where the non-time-to-event margin Y is binary, see Figure 2(a) for a plot of the non-commensurable margins analysed in Appendix B. However, as shown in Section 5.2 and Appendices E2 and E3, other combinations of bivariate outcomes can be accommodated using the proposed approach. The log-likelihood of the i-th observation in case of a binary non-time-to-event margin is then:

$$\begin{aligned} \ell = (1-y)\ln\{C[F_1(0), F_2(\boldsymbol{\delta}); \vartheta^{(c)}] - C[F_1(0), F_2(\boldsymbol{\delta}) - f_2(\boldsymbol{\delta}); \vartheta^{(c)}]\} + \\ y\ln\{f_2(\boldsymbol{\delta}) - C[F_1(0), F_2(\boldsymbol{\delta}); \vartheta^{(c)}] + C[F_1(0), F_2(\boldsymbol{\delta}) - f_2(\boldsymbol{\delta}); \vartheta^{(c)}]\}, \end{aligned}$$

where the copula $C(\cdot)$ is evaluated using the CDF of y at zero, i.e. $F_1(0)$ cf. Marra et al. (2020), and $F_2(\boldsymbol{\delta})$ is the proposed function based on DT or PW techniques.

A crucial aspect of copula modelling is the preservation of the dependence struc-

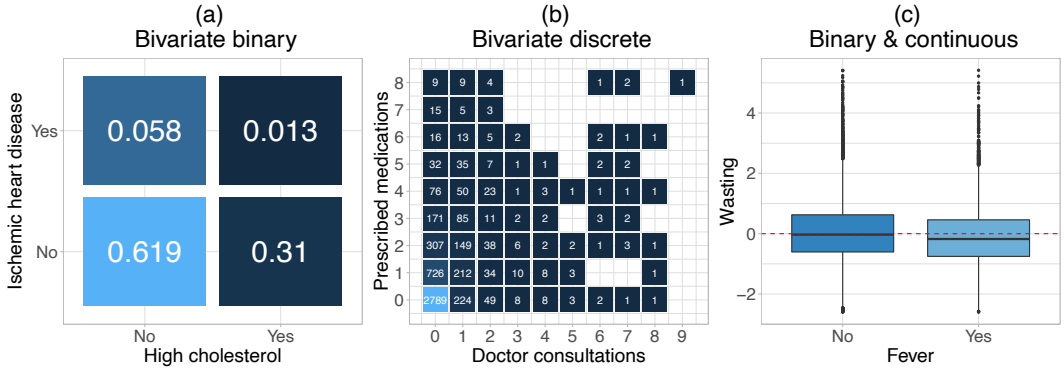

Figure 4: Examples of bivariate responses analysed in Appendix C: Bivariate binary (a), bivariate discrete (b) and mixed binary & continuous (c).

ture between the margins when applying a monotonic transformation to them (Nelsen, 2006; Klement et al., 2002). The proposed function $F(\boldsymbol{\delta})$ based on DT and PW approaches is a monotonic decreasing function of the original variable, we refer to Supplement A of Appendix B for a proof. Therefore, replacing the survival function of the time-to-event margin with the proposed function $F(\boldsymbol{\delta})$ in a bivariate copula model preserves the original dependence structure of interest.

2.4 Data-driven variable selection

Model-building encompasses aspects related to selection of the best-fitting distribution to the data, as well as the selection and allocation of the most informative variables to the different parameters of the assumed conditional response distribution. We now focus on the latter, i.e. variable selection and allocation. Some established techniques for variable selection are step-wise selection using information criteria (James et al., 2021), hypothesis-testing and removal of "non-significant" variables (Jenssen et al., 2002; Chowdhury and Turin, 2020), and the LASSO (Tibshirani, 2018). A wide range of methods exist for Bayesian inference as well, see for example Tadesse and Vannucci (2021).

As described in Section 2.1, distributional regression allows to model potentially all parameters of a uni- or multivariate distribution using a wide range of covariate effects in each additive predictor. However, as a trade-off to this great degree of flexibility, model-building and variable selection become increasingly difficult for univariate responses with a moderate number of covariates. If there are P potential covariates in the data, then a total of 2^P distinct subsets have to be fitted

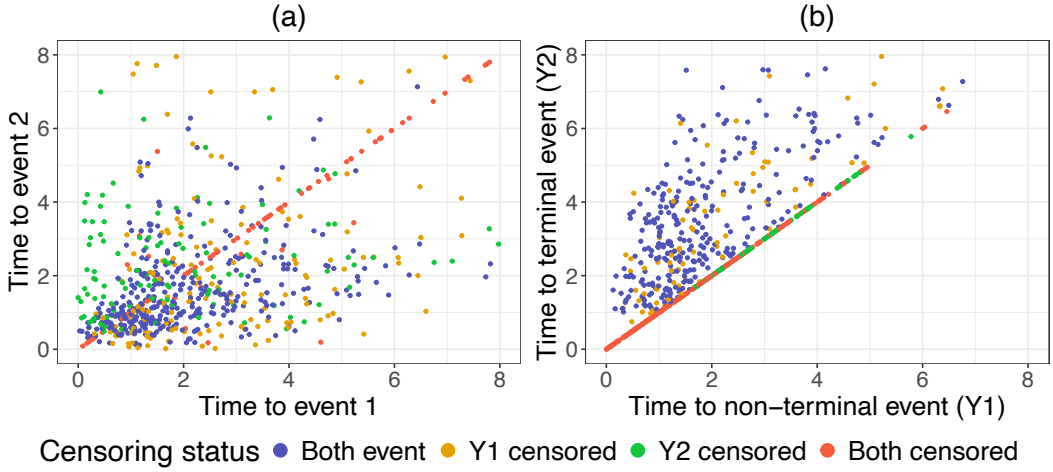

Figure 5: Illustration of synthetic bivariate time-to-event responses with right-censoring (a) and semi-competing risks (b).

in order to find an optimal model. This is already challenging when modelling distributions that consist of one parameter, e.g. Poisson or Bernoulli distributions, and might be even be infeasible for distributions with $K > 1$ parameters. The problem is further exacerbated when the number of responses increases in models such as those in Equation (2), due to the joint distribution tending to have an even larger number of parameters, since now $K = K_1 + K_2 + 1$, with K_1 and K_2 potentially being large. The last two articles included in this thesis aim to address this issue by proposing to estimate all coefficients of copula-based distributional regression models for various types of bivariate response structures via statistical boosting. The considered response types are bivariate binary, bivariate discrete, bivariate mixed binary & continuous, and bivariate time-to-event with right-censoring. Figure 4 shows examples of the former three response types, whereas Figure 5 shows examples of the latter. The estimation approach involves a component-wise gradient boosting algorithm with regression type base-learners (Friedman, 2001; Bühlmann and Hothorn, 2007).

Component-wise gradient boosting for distributional regression

Boosting originated in machine learning and has been extended towards estimating statistical models (Mayr et al., 2014). The base-learners correspond to the smooth functions of the covariates in the structured additive predictors from Equation (4), i.e. $s_{rk}^{(\bullet)}(x_{rk})$, $\bullet \in \{1, 2, c\}$. Typically the base-learners are based on a single regressor specific to the type of covariate effect. For example linear functions, P-splines for non-linear effects or Gaussian Markov Random Fields for

structured discrete spatial effects. Hothorn et al. (2010) and Mayr et al. (2012) provide a list of currently implemented base-learners available for the R packages `mboost` (Hothorn et al., 2022) and `gamboostLSS` (Hofner et al., 2016).

Boosting minimises the *empirical risk* $\omega_n = \frac{1}{n}\sum_{i=1}^{n}\omega(\boldsymbol{y}_i;\boldsymbol{\vartheta}_i)$ iteratively. Where $\boldsymbol{\vartheta}_i = \left(\boldsymbol{\vartheta}_i^{(1)}, \boldsymbol{\vartheta}_i^{(2)}, \vartheta_i^{(c)}\right) \in \mathbb{R}^K$ is the distribution parameter vector of the i-th observation, K denotes the number of parameters of the joint distribution, and $\omega(\cdot)$ represents the loss function of interest. For distributional regression models, the loss function corresponds to the negative log-likelihood, i.e. $\omega(\cdot) = -\ell_i(\cdot)$, but other loss functions can be used as well. In every iteration, the algorithm fits each of the pre-specified base-learners of each distribution parameter individually to the negative gradient of the loss w.r.t. to the additive predictors, i.e. $-\partial\omega(\boldsymbol{y}_i;\boldsymbol{\eta}_i)/\partial\eta_k^{(\bullet)}$, often called *pseudo-residuals*.

Our approach uses "non-cyclical updates" (Thomas et al., 2018), which means that only the best-fitting base-learner among all distribution parameters per fitting iteration is selected based on the potential decrease in the empirical risk. Then, a "weak" update of the model is conducted. Since the type of base-learner for each covariate is unmodified throughout the fitting process, the final covariate effect remains of the same type. The procedure is run for a pre-specified number of iterations denoted by $\mathtt{m}_{\mathtt{stop}}$. This quantity plays a similar role like the LASSO penalty parameter typically denoted by "λ" (Hepp et al., 2016), and acts as the main tuning parameter. By conducting *early stopping*, i.e. using $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}} < \mathtt{m}_{\mathtt{stop}}$ iterations, some base-learners will effectively be left out of the model due to not being selected in any of the iterations. This results in data-driven variable selection as well as shrinkage of covariate effects.

As a trade-off for the intrinsic data-driven variable selection mechanism, statistical boosting lacks "out-of-the-box" uncertainty quantification measures that one would obtain from penalised maximum likelihood or Bayesian inference such as standard errors of the estimated coefficients. However, it should be noted that data-driven variable selection and regularisation of effect estimates resulting from boosting with early stopping are particularly suitable for exploratory data analyses or prediction modelling. Boosting is able to provide valuable insights by automatically selecting relevant variables without requiring prior knowledge of their importance and may be even used in tandem with classical estimation techniques, see for example Gioia et al. (2022).

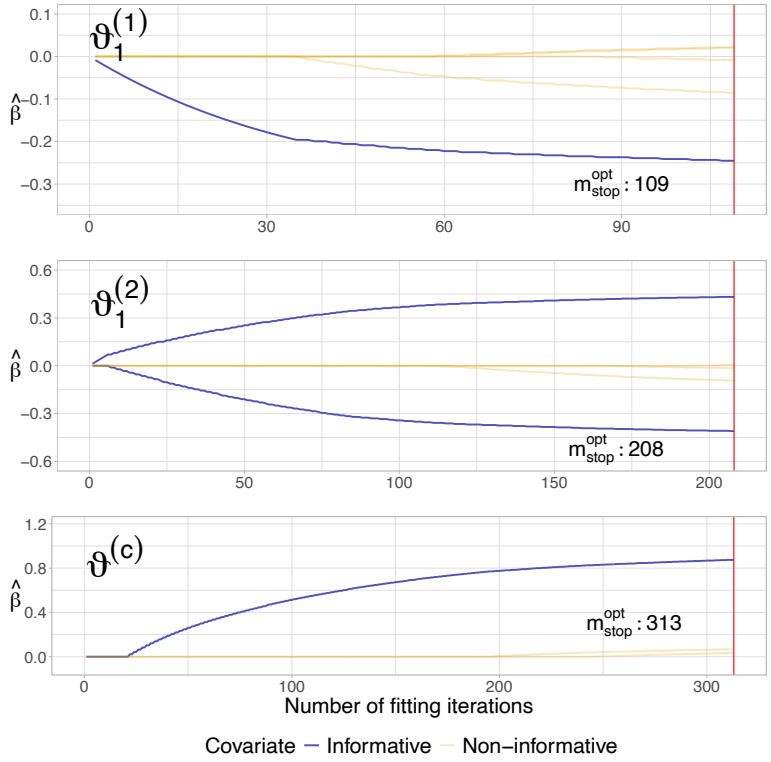

Figure 6: Illustration of the fitting iterations of boosting a bivariate distribution consisting of three parameters using synthetic data. Red vertical line indicates the optimal number of fitting iterations.

Implementational details

The proposed approach extends the boosting methodology of Hans et al. (2023) to accommodate bivariate binary, bivariate discrete, bivariate mixed binary & continuous, as well as bivariate right-censored time-to-event data. The boosting algorithm for the former three type of responses is presented in Appendix C, see Figure 4 for a visualisation of those response structures. The procedure for the latter response type is introduced in Appendix D. Figure 5 depicts an example of synthetic bivariate right-censored time-to-event data.

The boosting algorithm for bivariate binary, bivariate discrete and mixed binary & continuous responses estimates all coefficients of the distributional regression model simultaneously. The software constructs a loss function based on a copula with suitable marginal distributions selected by the analyst and estimates the coefficients of all model coefficients simultaneously. When the interest lies in the joint survival function of bivariate right-censored time-to-event responses, estimation is conducted in a two-step fashion akin to Joe (2005). First, the

sub-models of the right-censored margins are boosted separately. Afterwards, the marginal survival and density functions are plugged into the copula-based loss, which is boosted as a function of the dependence parameter $\vartheta^{(c)}$. Figure 6 illustrates the iterative fitting procedure of boosting using a synthetic bivariate binary response whose joint distribution consists of three parameters.

Both boosting algorithms have been integrated into the R package `gamboostLSS` (Hofner et al., 2016). They encompass a wide range of marginal distributions for discrete, continuous and right-censored time-to-event responses, as well as copula functions. Due to the modular structure, it is possible to further add user-specific marginal distributions and copulas. In tandem with the method of Hans et al. (2023), the proposed boosting algorithms allow for data-driven variable selection in distributional regression models for bivariate continuous, bivariate binary, bivariate discrete, bivariate right-censored time-to-event and mixed binary & continuous responses. We refer to Appendix C for details on the simultaneous estimation algorithm used for bivariate binary, discrete and mixed binary & continuous responses. Appendix E4 demonstrates how to fit statistical models via boosting to the aforementioned type of responses in `R` using `gamboostLSS`. Appendix D contains the details of the two-step algorithm used for bivariate right-censored outcomes. See Appendix E5 for illustrations on how to use the software implementation for boosting distributional regression models for bivariate time-to-event data.

3 Summaries of the contributed manuscripts

This section is comprised of short summaries of each of the four contributed manuscripts that constitute the dissertation. The contributed articles can be found in Appendices A, B, C, and D, respectively. In fulfilment of the requirements for a cumulative dissertation at the Department of Statistics of the TU Dortmund University (Fakultät Statistik der Technische Universität Dortmund), my contributions and those of all listed co-authors are declared in full detail.

Flexible instrumental variable distributional regression

This article is joint work with Maike Hohberg, Andreas Groll and Thomas Kneib. The article is published in the *Journal of the Royal Statistical Society Series A: Statistics in Society.* https://doi.org/10.1111/rssa.12598.

Briseño Sanchez, G., Hohberg, M., Groll, A., & Kneib, T. (2020). Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183 (4), 1553–1574.

The corresponding published article can be found in Appendix A.

This article is concerned with addressing two notable limitations of standard instrumental variable regression, a technique that is widely applied in experimental and observational studies: restricted estimation to the conditional mean of the outcome and the assumption of a linear relationship between regressors and outcome. The main contribution is to propose an instrumental variable estimation procedure built on distributional regression that combines moderns and flexible techniques with causal analysis. The proposed approach provides the means to model all parameters of potentially a complex response distributions by considering also nonlinear relationships between explanatory variables, instrument and outcome. In addition, the article introduces "distributional treatment effects", which generalise the notion of treatment effects to other characteristics or features of an estimated conditional response distribution. Estimates of changes in the expected value, variance, coefficient of variation, etc. may now be obtained from a fitted distributional regression model under the presence of an endogenous treatment variable.

In the article, an extensive simulation study is presented in which the performance of the proposed instrumental variable distributional regression method is investigated under different endogenous treatment, response types and sample sizes. Moreover, the performance of bootstrap confidence intervals is analysed. The simulations show that the proposed approach exhibits satisfactory performance at recovering the true treatment effect on the expected value as well as on the standard deviation of the response variable.

The proposed instrumental variable distributional regression method is applied in an empirical study to estimate the effect of a binary endogenous treatment variable (presence of rural electrification program) on female and male employment in the South African province of KwaZulu-Natal. The analysis found positive treatment effects on the mean for female employment rates, negative effects for male employment, and a negative effect on the standard deviation for both genders. These findings indicate a homogenisation in employment rates due to the rural electrification program. The application demonstrates the potentials of using distributional regression in tandem with instrumental variable regression to account for endogeneity and to estimate treatment effects beyond the mean. While applied in a labour economics context, instrumental variable distributional regression estimators could be used in other domains as well.

My contributions to the article are the following:

- Conceptualisation and writing of the article.
- Formalisation of the proposed model class and its implementation in R.
- Implementation of the models in R.
- Design, implementation and evaluation of the results of the simulation studies.
- Implementation and evaluation of the application.
- Illustration of the results of the simulation studies and the application graphically using the R package `ggplot2`.
- Writing of the entirety of the supplementary material.

Maike Hohberg contributed to the conceptualisation, framing and writing of the article, as well as the formalisation of the proposed model class. Maike Hohberg prepared the data analysed in the empirical study and contributed to

the evaluation and interpretation of the results from said analysis. Thomas Kneib and Andreas Groll advised on methodological and technical challenges. All listed authors revised and edited the final version of the article.

Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses

This article is joint work with Andreas Groll.

Briseño Sanchez, G. & Groll, A. (2024) Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses. [Manuscript submitted for publication.]

The corresponding article can be found in Appendix B.

In this article, a copula-based distributional regression modelling approach for bivariate responses comprised of non-commensurate (i.e. mixed) variables is proposed. The bivariate mixed response consists of a right-censored time-to-event outcome and a non-time-to-event variable. The underlying hazard rate of the time-to-event margin is modelled using discrete-time-to-event or piecewise-exponential methods. A flexible statistical model is achieved by relying on the correspondence of the likelihood of the aforementioned time-to-event approaches with univariate Bernoulli and Poisson distributions, respectively. The joint bivariate distribution of such mixed responses is constructed by means of parametric bivariate copulas. This allows for separate specification of the dependence structure between the margins and their individual distribution functions. A mathematical proof shows that using the proposed functions based on the likelihood of a piecewise-exponential or discrete time-to-event model results in a monotonic function of the original event time outcome. Therefore, the initial dependence structure of interest is preserved when using the proposed approach. All of the model coefficients are estimated simultaneously via penalised maximum likelihood. The proposed model class is implemented within a modified version of the R package `GJRM` (Marra and Radice, 2023).

Extensive simulation studies were conducted in order to assess the model fit and suitability of the likelihood used to model the time-to-event margin within a copula regression context. Results from the simulations indicate a good performance in terms of coefficient estimation in both marginal responses and the

dependence parameter and estimation of the hazard rate in different censoring regimes. In order to showcase the versatility of the proposed model for mixed responses, an analysis of data on the red-light running behaviour of E-cyclists is presented. The mixed response consists of a binary response that indicates whether a cyclist checks for incoming traffic while waiting at a road junction, and a time-to-event variable that indicates the time at which the cyclist crossed a street while the traffic light showed red. Two model configurations are fitted to the mixed response. The first one models the hazard rate of the event time variable using a piecewise-exponential approach, and another with discrete time-to-event techniques. The results of the analysis show that both approaches yield very similar results. Both find a moderate dependence between the binary and the time-to-event responses. Additionally, takeaway delivery E-cyclists engage in riskier driving behaviours compared to those that are not employed in such occupations.

My contributions to the article are the following:

- Conceptualisation of the proposed model and derivation of the explicit mathematical definitions of the functions postulated in the article, as well as their respective first and second order partial derivatives.
- Formalisation of the proof provided in supplementary material A of the contributed article.
- Implementation of the proposed model in the R package `GJRM` (version 0.2-5). This comprises the postulated functions their respective first and second order partial derivatives, as well as various necessary modifications to allow `GJRM` to run with these new configurations.
- Design, implementation and evaluation of the results of the simulation studies.
- Statistical analysis of the data presented in Section 4 of the article.
- Illustration of the results of the simulation studies and the application graphically using the R package `ggplot2`.
- Conceptualisation, framing and writing of the manuscript as well as all of its supplementary material.

All listed authors contributed to the conceptualisation and framing of the article. Andreas Groll also contributed to early discussions that led to the explicit

definitions of the main contributions presented in the article. An early version of the article submitted to the 34-th International Workshop on Statistical Modelling held at the University of Trieste was awarded the "best student paper" (Briseño Sanchez and Groll, 2022). Andreas Groll verified the proof provided in supplementary material A of the contributed article. All listed authors revised and edited the final version of the article.

Boosting distributional copula regression for bivariate binary, discrete and mixed responses

This article is joint work with Nadja Klein, Hannah Klinkhammer, and Andreas Mayr. The article is published in the journal *Statistical Methods in Medical Research*, https://doi.org/10.1177/09622802241313294.

Briseño Sanchez, G., Klein, N., Klinkhammer, H., & Mayr, A. (2025). Boosting distributional copula regression for bivariate binary, discrete and mixed responses. *Statistical Methods in Medical Research*, https://doi.org/10.1177/09622802241313294.

The corresponding article can be found in Appendix C.

The article develops statistical boosting for distributional regression based on bivariate copulas with arbitrary marginal distributions. Such models are suited to analyse bivariate binary, count, continuous or bivariate non-commensurate (i.e. mixed) outcomes. The article was motivated by the rather high prevalence of binary and count response analyses conducted in medical research. Efficient and scalable estimation is proposed by means of an adapted component-wise gradient boosting algorithm with statistical models as base-learners. As opposed to classical likelihood or Bayesian estimation, a key benefit of boosting is an implicit data-driven variable selection mechanism as well as shrinkage without additional input or assumptions from the analyst. To the best of our knowledge, the implementation presented in the article is the only one that combines a wide range of covariate effects, marginal distributions, copula functions, and implicit data-driven variable selection.

Various simulation studies were conducted in order to assess the performance of the proposed boosting algorithm. In particular, the variable selection accu-

racy, predictive and probabilistic performance were investigated against models with independent margins in settings with increasing number of covariates and three different type of bivariate responses: bivariate binary, bivariate count and mixed binary & continuous responses. All cases considered different dependence structures with varying dependence strength. Results from the simulations show that the proposed boosted distributional regression approach performs well in all aforementioned scenarios.

Motivated by challenges in the analysis of biomedical data, the versatility of the approach is showcased on data from genetic epidemiology, healthcare utilisation and childhood undernutrition. These case studies show that the combination of distributional regression and statistical boosting leads to a streamlined model-building process. All of the developments presented in the article are implemented in the R package `gamboostLSS`.

My contributions to the article are the following:

- Conceptualisation and framing of the article.
- Implementation of the proposed model class within the previously established simultaneous estimation framework as add-on functions for the R package `gamboostLSS`.
- Extending the catalogue of implemented distributions by adding univariate discrete, binary and some additional continuous distributions.
- Design, implementation and evaluation of the results of the simulation studies.
- Illustration of the results of the simulation studies and the biomedical applications graphically using the R package `ggplot2`.
- Writing the majority of the manuscript and all of the supplementary material.

Nadja Klein and Andreas Mayr contributed to the conceptualisation and framing of the project. Hannah Klinkhammer provided the UK Biobank data (under application number 81202) analysed in Section 4.1 of the article, as well as the R code used to create the Manhattan plot that depicts the results of the fitted model (Figure 2 in the article). Nadja Klein provided helpful guidance to analyse the Demographic and Health Surveys (DHS; Demographic and Health Survey, 2023) data from Section 4.3 of the article. All listed authors revised and edited

the final version of the article.

Boosting distributional copula regression for bivariate right-censored time-to-event data

This article is joint work with Nadja Klein, Andreas Groll, and Andreas Mayr. The article is currently published on arXiv: https://doi.org/10.48550/arXiv.2412.15041.

Briseño Sanchez, G., Klein, N., Groll, A., & Mayr, A. (2024). Boosting distributional copula regression for bivariate right-censored time-to-event data. [Manuscript submitted for publication.]

The corresponding article can be found in Appendix D.

The final article extends distributional regression with intrinsic, data-driven variable selection to bivariate time-to-event data structures in the presence of independent right-censoring. The statistical models of the marginal right-censored time-to-event responses are specified using well-known parametric distributions such as the log-Normal, log-Logistic (proportional odds model), or Weibull (proportional hazards model) distributions. Such models are typically known in the literature as parametric or *Accelerated Failure Time* (AFT) models (Klein and Moeschberger, 2003; Kalbfleisch and Prentice, 2002). The bivariate joint survival function is constructed using parametric copulas, allowing for a separate specification of the dependence structure between the time-to-event outcome variables and their respective marginal survival distributions.

A two-step component-wise gradient-based boosting algorithm is devised to estimate the model coefficients. In the first step, the margins are boosted independently. The fitted values of the marginal survival and density functions are plugged-in to the log-likelihood function. In the second step, the log-likelihood of the model is optimised as a function of the additive predictor of the copula dependence parameter. The proposed boosting approach is able to conduct data-driven variable selection, a feature that is extremely helpful in such a complex model class since the specified joint survival functions are likely to involve a large number of parameters. To the best of our knowledge, this is the first implementation of multivariate AFT models via distributional regression with automatic

variable selection by means of statistical boosting. In addition, the proposed approach works in high-dimensional ($p \gg n$) settings, where estimation of the proposed model class using classical techniques would be infeasible.

The article features extensive simulation studies where the selection of informative covariates, predictive performance and copula selection capabilities are investigated. Classic bivariate right-censored time-to-event as well as semi-competing risks responses are considered throughout the article. The simulations show that the proposed method outperforms independent models and other classical time-to-event analysis techniques like the Cox proportional hazards model. This holds for increasing number of candidate covariates in the model as well as different censoring regimes. All compared approaches were fitted using boosting.

The practical potential of the proposed model class is illustrated on a high-dimensional application related to semi-competing risks responses in ovarian cancer. The analysed data stemmed from the Bioconductor `curatedOvarianData` package. The joint survival function of time of tumour-progression and time of death is modelled as function of covariates. Over 11,000 gene expressions and two clinical variables are included in the data. The proposed method determined without further input from the analyst that around 100 covariates were informative for the bivariate joint survival function. All of the developments presented in the article were implemented in the R package `gamboostLSS`.

My contributions to the article are the following:

- Conceptualisation and framing of the article.
- Formalisation and implementation of the proposed model class in its modified two-step procedure as add-on functions for the R package `gamboostLSS`.
- Design, implementation and evaluation of the results of the simulation studies.
- Statistical analysis of the data presented in Section 4 of the article.
- Illustration of the results of the simulation studies and the application graphically using the R package `ggplot2`.
- Writing of the majority of the manuscript and all of the supplementary material.

All listed authors contributed to the conceptualisation and framing of the article.

Nadja Klein and Andreas Mayr contributed to early discussions that led to the implementation of the two-step algorithm. Moreover, Nadja Klein and Andreas Mayr provided helpful suggestions to the design of the simulation studies as well as the presentation of their results. All listed authors revised and edited the final version of the article.

4 Concluding remarks

This cumulative dissertation provides four articles on distributional regression applied to causal inference, modelling of bivariate non-commensurable responses and data-driven variable selection for copula-based models of the aforementioned class. In addition, it develops software to further extend the applicability of the proposed advancements via comprehensible and reusable tools.

The first article (Appendix A) introduced 2SGAMLSS, an instrumental variable distributional regression estimator used to provide a wide range of treatment effects on various distributional features of an outcome. The proposed method addresses model-fitting under endogeneity due to unmeasured confounders and various treatment types, such as binary and complex treatments expressed as continuous variables. The 2SGAMLSS method may be fitted using any software library for distributional regression.

In the second article (Appendix B), the joint distribution of mixed responses with one right-censored time-to-event component is modelled using functions derived from piecewise-exponential and discrete-time-to-event likelihoods. This approach embeds flexible semi-parametric regression models for time-to-event variables into a copula-based distributional regression framework combined with simultaneous estimation of all model coefficients using penalised maximum likelihood. The proposed model class is implemented in a modified version of the R package `GJRM`, allowing for convenient access of the implemented methodology for empirical analyses.

The last two articles are dedicated to develop intrinsic data-driven variable selection methods for copula-based distributional regression models for various response structures. The unknown model coefficients are estimated using statistical boosting, which has the additional advantage of accommodating high-dimensional covariates. On one hand, the third article (Appendix C) is concerned with accommodating bivariate binary, count and non-commensurate responses that consist of a binary and a continuous component. On the other hand, the fourth article (Appendix D) specialises in bivariate right-censored time-to-event outcomes. The aforementioned boosting algorithms have been implemented as supplements for the R package `gamboostLSS`.

The use of univariate and copula-based distributional regression for analysing challenging data structures that feature complex responses is shown to be ben-

eficial. These sophisticated modelling techniques may help uncover previously unknown associations between the covariates and the response, as well as dependencies between the analysed responses. All of these aspects make it possible to derive more informed conclusions from the data. The application of data-driven variable selection techniques based on statistical boosting shows how more complex distributional models can be easily fitted to complex data structures accompanied by potentially high-dimensional covariates in a streamlined fashion, even in scenarios without *a-priori* knowledge of the most influential factors.

Lastly, the computation of uncertainty quantification measures remains a caveat of the methodology proposed in the first, third and fourth contributed articles. While in the first case adopting a copula-based modelling approach as described in Section 5.1 mitigates this issue, it remains largely unaddressed for statistical boosting.

5 Outlook

During the development of this dissertation further advancements of the presented topics were discussed and explored to a varying degree of depth. This section aims to provide a short discussion on possible avenues of future research.

5.1 Refinements of distributional regression models under endogeneity due to confounding

The proposed 2SGAMLSS is able to accurately recover treatment effects on various distributional features of a response variable in the presence of an endogenous treatment variable. However, 2SGAMLSS suffers from two main drawbacks: The first is the lack of a data-driven variable selection mechanism. The second is cumbersome computation of uncertainty quantification measures such as standard errors of the estimated coefficients. This is due to a bootstrap inference procedure that requires re-fitting the second-stage multiple times which can be computationally expensive. Below we discuss some ideas on how to tackle these pitfalls. The first drawback may be addressed using statistical boosting, whereas for the second one we propose to rely on copula functions.

Statistical boosting in causal inference using 2SGAMLSS

Briseño Sanchez and Groll (2020) briefly explore the use of statistical boosting in tandem with 2SGAMLSS to estimate all unknown model coefficients and conduct data-driven variable selection under endogeneity. The analysis in Briseño Sanchez and Groll (2020) replicates that of Briseño Sanchez et al. (2020) (Appendix E), but each stage of 2SGAMLSS is estimated using statistical boosting via the package `gamboostLSS`. The results from both analyses agree regarding the causal effects. Most importantly, the boosted 2SGAMLSS provided a more parsimonious model compared to its penalised maximum likelihood counterpart.

The combined use of statistical boosting and 2SGAMLSS for causal inference can be beneficial in analyses with a moderate to large number of covariates or when there is no strong (or any) evidence of which covariates affect treatment and response. Below we show what we consider to be the most promising avenue of future development for 2SGAMLSS, however adding statistical boosting to the mix of estimation methods for distributional regression in causal inference should be further investigated via extensive simulations, re-analysis of previously

published data and perhaps most importantly a comparative study of regression methods for endogeneity due to unmeasured confounders.

Addressing endogeneity using copulas

Instead of using a two-stage estimation approach, a potential avenue of improvement for 2SGAMLSS would be to construct the joint distribution of the endogenous treatment and the response using a copula function:

$$F(y_{1i},\ y_{2i};\ \boldsymbol{\vartheta}_i) = C[F_1(y_{1i};\ \boldsymbol{\vartheta}_i^{(1)}),\ F_2(y_{2i};\ \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}],$$

where in this case y_{1i} represents the endogenous treatment variable, denoted by x_{en} in Subsection 2.2, and y_{2i} is the outcome variable, see e.g. Wyszynski and Marra (2018); Marra et al. (2020); Wiemann et al. (2022), Marra et al. (2023) and Marra and Radice (2024). The structured additive predictors corresponding to the first-stage model from Equation (5) belong to the first margin, i.e. $\boldsymbol{\vartheta}_i^{(1)}$, whereas those of the second stage from Equation (7) correspond to the second margin ($\boldsymbol{\vartheta}_i^{(2)}$). The copula-based model does not require the estimation of the first-stage residuals $\hat{\xi}$ and they would therefore be omitted from the additive predictors belonging to the response. One would be able to estimate all model coefficients simultaneously and account for the endogeneity via the dependence parameter $\vartheta_i^{(c)}$ (Han and Vytlacil, 2017). Additional benefits include a gain in efficiency, and the removal of bootstrap inference for 2SGAMLSS.

Note that constructing the joint distribution of the endogenous variable and the response using copulas leads to joint distributions of mixed responses whenever treatment and outcome variables are non-commensurable, e.g. binary treatment and continuous response. As a trade-off, a bivariate copula approach would limit the analysis to one endogenous treatment variable. Models for three-dimensional responses considering two endogenous binary variables and a binary outcome have been developed in Filippou et al. (2022), and Marra et al. (2024) model two responses and one endogenous variable in a switching regimes framework. However, model complexity increases drastically when considering multivariate distributions. The approach proposed by Kock and Klein (2024) for multivariate responses of arbitrary dimensions via the Gaussian copula might be a viable alternative in cases with multiple endogenous variables or multiple responses, albeit their work would need to be adapted to accommodate recursive model structures, i.e. when one marginal response appears in the model of the other.

5.2 Modelling of other types of mixed non-time-to-event & time-to-event responses

The proposed models for mixed responses with time-to-event margins based on Bernoulli and Poisson likelihoods can be used for different combinations of mixed responses. We present three further applications of the proposed approaches: discrete & time-to-event, continuous & time-to-event and the special case of bivariate right-censored time-to-event responses. For the remainder of this subsection the time-to-event response corresponds to the second margin and its survival function is replaced by the proposed function $F(\boldsymbol{\delta}_i)$ as in Section 2.3. In cases where the non-time-to-event margin is discrete, the log-likelihood for the bivariate mixed response is:

$$\begin{aligned}\ell_i = \ln\{&C[F_1(y_{1i}), F_2(\boldsymbol{\delta}_i)] - C[F_1(y_{1i}) - f_1(y_{1i}), F_2(\boldsymbol{\delta}_i)] - \\ &C[F_1(y_{1i}), F_2(\boldsymbol{\delta}_i) - f_2(\boldsymbol{\delta}_i)] + C[F_1(y_{1i}) - f_1(y_{1i}), F_2(\boldsymbol{\delta}_i) - f_2(\boldsymbol{\delta}_i)]\}, \quad (12)\end{aligned}$$

where $F_1(\cdot)$ is the CDF of a discrete response, for example, a Poisson or a negative Binomial distribution, cf. van der Wurp et al. (2020). For cases when the non-time-to-event margin is continuous, the log-likelihood for a bivariate mixed discrete and continuous response is:

$$\ell_i = \ln\left\{\frac{\partial C[F_1(y_{1i}), F_2(\boldsymbol{\delta}_i)]}{\partial F_1(y_{1i})} - \frac{\partial C[F_1(y_{1i}), F_2(\boldsymbol{\delta}_i) - f_2(\boldsymbol{\delta}_i)]}{\partial F_1(y_{1i})}\right\} + \ln[f_1(y_{1i})],$$

where $F_1(\cdot)$ is the CDF of a continuous response that may depend on multiple parameters, for example, a Gaussian distribution with 2 parameters, or a Dagum distribution with 3 parameters. Note that the parameter vectors of the margins $(\boldsymbol{\vartheta}_i^{(1)}, \boldsymbol{\vartheta}_i^{(2)})$ and the copula dependence parameter $\vartheta_i^{(c)}$ have been omitted to avoid clutter in the notation. We conducted a simulation study in order to assess the performance of the proposed functions based on *DT* and *PW* likelihoods. We refer to Appendix E2 for illustrations on how to use a modified version of `GJRM` to fit models to these responses and Appendix F for more details on the simulation study.

Bivariate right-censored time-to-event data using $F(\boldsymbol{\delta})$

Bivariate time-to-event variables may be analysed by replacing the marginal survival functions with the respective proposed functions $F_1(\boldsymbol{\delta}_1)$, $F_2(\boldsymbol{\delta}_2)$ based on *DT* or *PW* techniques. Here $\boldsymbol{\delta}_1$ and $\boldsymbol{\delta}_2$ denote the respective sequences of

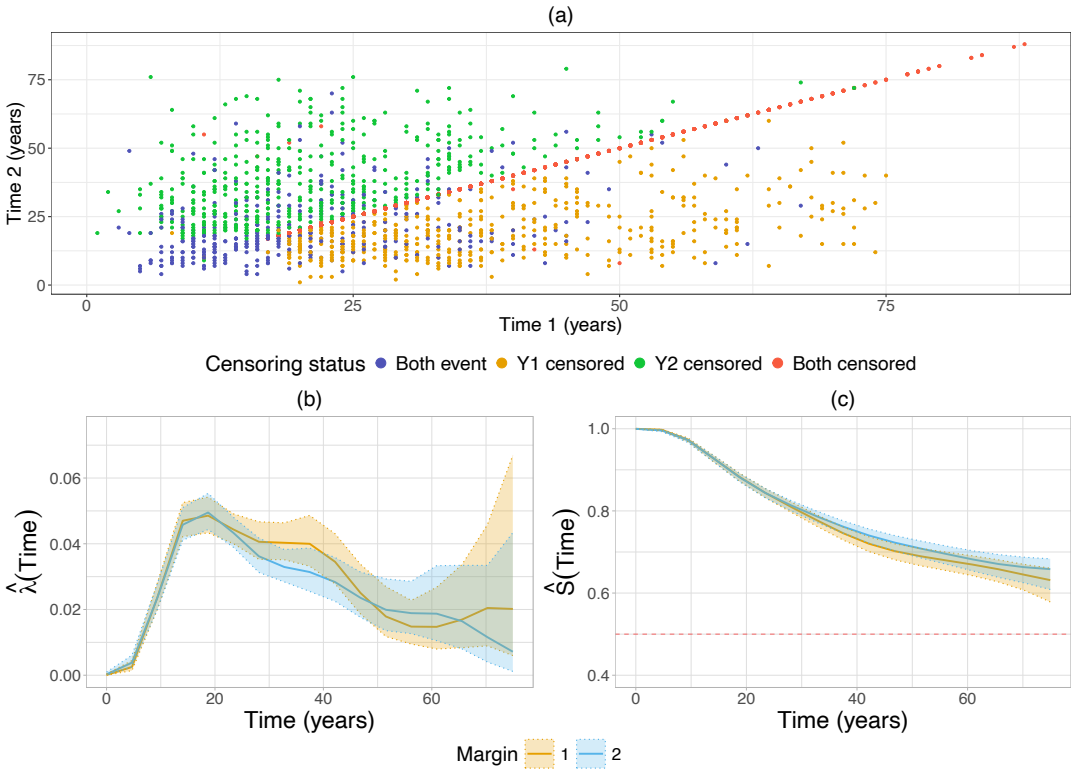

Figure 7: Scatterplot of the Australian twins data (a). Estimated baseline marginal hazard probabilities (b) and baseline marginal survival functions (c) with point-wise 95% confidence intervals. Red dashed line indicates the median.

auxiliary variables obtained from conducting the data augmentation procedure described in Section 2.3. A copula-based distributional regression model may be fitted by plugging $F_1(\boldsymbol{\delta}_1)$ and $F_2(\boldsymbol{\delta}_2)$ into the log-likelihood shown in Equation (12). See Appendix E3 for examples on how to fit distributional regression models for this type of responses using our modified version of `GJRM`.

We present a brief illustration using data on the time to appendectomy for adult Australian twins from Duffy et al. (1990), previously analysed by Romeo et al. (2018) and Marra and Radice (2020). The study investigated whether the strength of the dependence is different for monozygotic (MZ) and dizygotic (DZ) twin pairs with respect to the risk of the onset of acute appendicitis. A stronger dependence for MZ twin pairs would be indicative of a genetic effect on the risk of acute appendicitis, posing evidence for heredity in the onset of the disease (Romeo et al., 2018). The marginal event times are the age at appendectomy (or censoring) of each twin in the pair. The data consists of $n = 3808$ pair of twins and the censoring rates are at 77% and 77.9% for each margin. We refer to Figure 7(a) for a scatterplot of the time-to-event variables.

The covariates in the data are the zygotic type of the pair (MZ or DZ) and the gender combination of both twins in the pair (both male, both female or mixed). For this illustration, we opt for a *DT* model and conduct the data augmentation procedure using 20 equidistant intervals between $t = 0$ and $t = 89$ years. The copula model consists of the following additive predictors:

$$\eta_{ij}^{(1)} = s_0^{(1)}(j), \quad \eta_{ij}^{(2)} = s_0^{(2)}(j), \quad \eta_i^{(c)} = \beta_0^{(c)} + \beta_1^{(c)}\texttt{zygosity}_i + \beta_2^{(c)}\texttt{gender}_i,$$

where $s_0^{(\bullet)}(j)$, $\bullet = 1, 2$ are the baseline hazard probabilities at time interval j modelled using P-splines. We follow Marra and Radice (2020) and fit a Student's t copula with 3 degrees of freedom.

The results of our copula model using the *DT* approach agree with those of Romeo et al. (2018) and Marra and Radice (2020). We find that the dependence between the margins expressed as Kendall's τ with 95% confidence interval for MZ twins is $\hat{\tau}_{MZ} = 0.275$ $[0.217; 0.329]$, whereas the dependence for DZ twins is $\hat{\tau}_{DZ} = 0.177$ $[0.103; 0.249]$. Figure 7(b) shows the estimated baseline marginal hazard probabilities, whereas Figure 7(c) depicts the estimated baseline marginal survival functions. It can be seen that both estimated marginal functions follow very similar trends for each twin. Lastly, we found that the variable `gender` had no significant effect on the additive predictor of the dependence parameter. The estimated coefficients with their respective 95% confidence intervals were $\beta^{(c)}_{2,\texttt{gender:male}} = -0.118$ $[-0.266;\ 0.032]$ and $\beta^{(c)}_{2,\texttt{gender:mixed}} = -0.134$ $[-0.305;\ 0.032]$.

5.3 Data-driven variable selection for further response structures

In this subsection, we discuss potential instances to apply data-driven variable selection via statistical boosting by extending the algorithms for copula-based distributional regression models. The statistical boosting algorithms introduced in Section 2.4, and Appendices C and D have been adapted to accommodate the following types of response structures: bivariate mixed binary & discrete, bivariate mixed discrete & continuous and bivariate time-to-event with general censoring schemes.

Bivariate mixed binary & discrete responses

For this illustration, we consider one of the datasets analysed in Section 4.2

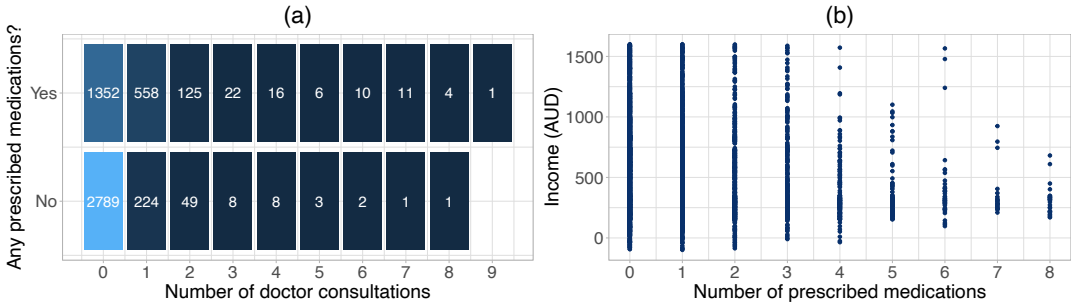

Figure 8: Scatterplot of bivariate mixed binary & discrete responses (a) as well as bivariate mixed discrete & continuous responses (b). Numbers inside the tiles represent the number of observations with that response combination.

of Appendix C related to doctor consultations and prescribed medications of $n = 5{,}190$ Australian healthcare recipients. We consider the discrete variable `prescrib` $\in \mathbb{N}$, which indicates the total number of prescribed medications used in the last 2 days, and dichotomise it into a binary variable using the following rule: `anyprescrib` $= \mathbb{1}\{$`prescrib` $> 0\}$. This produces the binary response `anyprescrib` $\in \{0, 1\}$, which indicates whether the i-th individual used one or more prescribed medications in the last two days. The bivariate mixed binary & discrete response is then comprised of `anyprescrib` $\in \{0, 1\}$ and `doctorco` $\in \mathbb{N}$, where `doctorco` indicates the number of consultations with a doctor or specialist in the past two weeks, see Figure 8(a) for a scatterplot of the non-commensurable margins. For this illustration we include the following two continuous and one binary covariates in the model: `age` (age in years divided by 100), `income` (annual income in Australian dollars divided by 1000), and `gender` (1 female, 0 male). The binary margin `anyprescrib` is modelled using a Bernoulli distribution with parameter $\vartheta_1^{(1)}$, whereas `doctorco` is modelled using a zero-altered logarithmic (ZALG) distribution with parameters $\vartheta_1^{(2)}$, and $\vartheta_2^{(2)}$.

The joint distribution of `anyprescrib` and `doctorco` is constructed using a Clayton copula. The log-likelihood function of such a mixed binary & discrete response is given by:

$$\begin{aligned}\ell_i = (1 - y_{1i}) \ln \left\{ C\left[F_1(0), F_2(y_{2i})\right] - C\left[F_1(0), F_2(y_{2i}) - f_2(y_{2i})\right] \right\} + \\ y_{1i} \ln \left\{ f_2(y_{2i}) - C\left[F_1(0), F_2(y_{2i})\right] + C\left[F_1(0), F_2(y_{2i}) - f_2(y_{2i})\right] \right\},\end{aligned}$$

where the parameter vector of the discrete margin $\boldsymbol{\vartheta}_i^{(2)}$ as well as the scalar parameters $\vartheta_i^{(1)}$ and $\vartheta_i^{(c)}$ have been omitted to avoid clutter, cf. Marra et al. (2020). The term $F_1(0)$ denotes the marginal distribution of the binary response

Table 2: Estimated linear effects of the models for bivariate mixed binary & discrete (a) and bivariate mixed discrete & continuous (b).

(a) Bivariate mixed binary & discrete

	Bernoulli	ZALG		Clayton Copula
	$\vartheta_1^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta^{(c)}$
	probit	logit	logit	log
`Intercept`	−0.521	−1.150	0.040	0.396
`gender` (female)	0.645	0.051	−0.170	−0.461

(b) Bivariate mixed discrete & continuous

	ZINBI			Log-logistic		Gaussian Copula
	$\vartheta_1^{(1)}$	$\vartheta_2^{(1)}$	$\vartheta_3^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta^{(c)}$
	log	log	logit	log	log	$\tanh^{-1}$
`Intercept`	−0.310	0.276	0.367	6.426	−2.327	0.013
`gender` (female)	0.256	−0.539	−1.634	−0.140	0.061	−0.024

Number of observations used for fitting: $n_{\text{train}} = 3892$.
Number of observations used for tuning $\texttt{m}_{\texttt{stop}}$: $n_{\texttt{mstop}} = 1298$.

evaluated at zero. The developed implementation for `gamboostLSS` can be used to fit a distributional regression model to such data. We refer to Appendix E4 for a brief demonstration on how to use the software.

Results of analysis of mixed binary & discrete response

Table 2(a) shows the estimated coefficients, whereas Figure 9 depicts the estimated non-linear effects across the four parameters of the joint bivariate distribution. The estimated Kendall's τ of the observations in the data lies within $\hat{\tau} = [0.279;\ 0.541]$, suggesting a moderate to strong dependence between the non-commensurable margins. The estimated linear effect of `gender` indicates that female individuals are more likely to use any number of prescribed medications compared to male individuals. In addition, the negative estimated coefficient on $\vartheta_2^{(2)}$ suggests that a higher number of doctor consultations can be expected from females, compared to males. This is because the parameter $\vartheta_2^{(2)}$ of the ZALG distribution directly models the probability of observing a zero. Moreover, the dependence between `anyprescrib` and `doctorco` is lower for female individuals, compared to male individuals.

In Figure 9(a) it can be seen that the estimated non-linear effect of `age` shows an increasing chance of visiting a doctor for older individuals. Such a trend makes

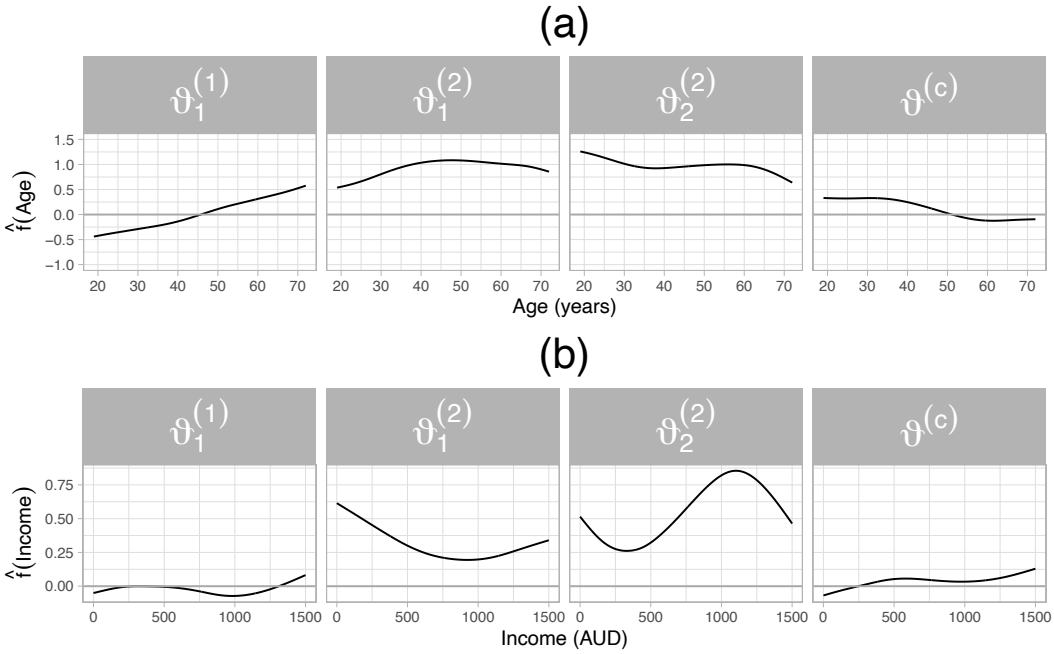

Figure 9: Estimated non-linear effects of `age` (a) and `income` (b) across the parameters of the joint bivariate distribution of `anydoctorco` and `prescrib`.

sense, since older individuals are more likely to visit a medical practitioner compared to younger people. Figure 9(b) shows that on one hand, the magnitude of the estimated effect of `income` is rather small on the parameter of `anyprescrib` and the copula dependence parameter $\vartheta^{(c)}$. On the other hand, `income` appears to have a more complex effect on the parameters of the distribution of `doctorco`. Lastly, the variable `age` has a downward-slopping estimated effect on the dependence parameter, suggesting a decreasing dependence between the margins for older individuals. The variable `income` exhibits an upward-sloping estimated effect on $\vartheta^{(c)}$, suggesting a slightly stronger dependence between the margins for individuals belonging to higher income brackets.

Bivariate mixed discrete & continuous responses

In this example we consider a bivariate non-commensurable response comprised of discrete and continuous margins, i.e. $Y_1 \in \mathbb{N}$ and $Y_2 \in \mathbb{R}$. The discrete margin is given by `prescrib` $\in \mathbb{N}$, whereas the continuous positive margin is given by the variable `income` $\in \mathbb{R}_+$, see Figure 8(b) for a scatterplot. For this illustration we consider the covariates `age` (age in years divided by 100), and `gender` (1 female, 0 male). In addition, we include `doctorco` as an explanatory variable. We model `prescrib` using a zero-inflated negative binomial (ZINBI)

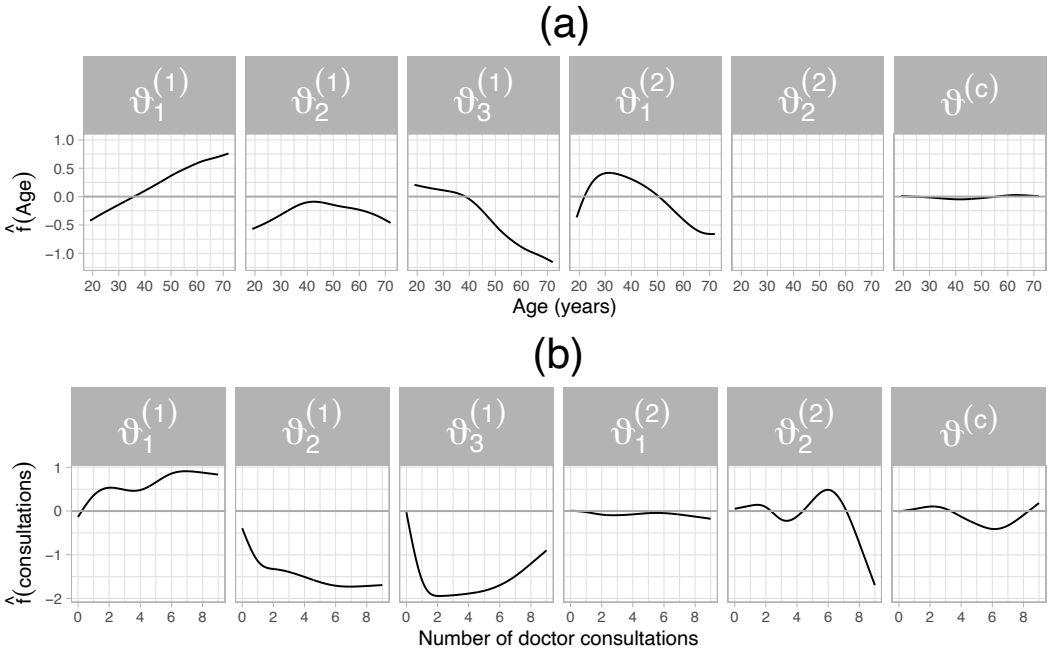

Figure 10: Estimated non-linear effects of `age` (a) and `doctorco` (b) across the parameters of the joint bivariate distribution of `prescrib` and `income`.

distribution with parameters $\vartheta_1^{(1)}$, $\vartheta_2^{(1)}$ and $\vartheta_3^{(1)}$, whereas `income` is modelled using a log-logistic distribution with parameters $\vartheta_1^{(2)}$ and $\vartheta_2^{(2)}$. In this illustration, the joint bivariate distribution is constructed using a Gaussian copula.

The log-likelihood function for bivariate non-commensurable discrete & continuous responses is then:

$$\ell_i = \ln \left\{ \frac{\partial C\left[F_1(y_{1i}), F_2(y_{2i})\right]}{\partial F_2(y_{2i})} - \frac{\partial C\left[F_1(y_{1i}) - f_1(y_{1i}, F_2(y_{2i}))\right]}{\partial F_2(y_{2i})} \right\} + \ln[f_2(y_{2i})],$$

where $f_1(y_{1i})$ denotes the PDF of the discrete margin, and once again the parameter vectors $\boldsymbol{\vartheta}_i^{(1)}, \boldsymbol{\vartheta}_i^{(2)}$ and the copula dependence parameter $\vartheta_i^{(c)}$ have been omitted to avoid clutter. Appendix E4 demonstrates how to fit a copula-based distributional regression model for this type of mixed responses using `gamboostLSS`.

Results of analysis of mixed discrete & continuous response

The estimated coefficients are shown in Table 2(b), whereas the estimated non-linear effects across the six parameters of the joint bivariate distribution are shown in Figure 10. In this example, the estimated Kendall's τ of the observations in the data lies within $\hat{\tau} = [-0.282;\ 0.141]$. This result shows that the dependence between the margins exhibits differences in direction as well as

magnitude for the different individuals in the data. The estimated linear effect of `gender` from Table 2(b) indicates that female individuals have a higher expected number of prescribed medications relative to males. The estimated effect of `gender` on the additive predictor of $\vartheta_3^{(1)}$ further supports this claim, since the parameter $\vartheta_3^{(1)}$ of the ZINBI distribution models the probability of observing a zero. Figure 10(a) shows the estimated non-linear effects of `age` on the additive predictors of the parameters of the bivariate distribution. An upward-sloping trend can be seen on $\vartheta_1^{(1)}$ can be seen, which points towards a higher expected number of used prescribed medications for older individuals. The downward-sloping estimated non-linear function on $\vartheta_3^{(1)}$ indicates a steadily decreasing chance of observing a zero as individuals become older. We remark that `age` was not selected for the parameter $\vartheta_2^{(2)}$ and has an effect with very small magnitude on $\vartheta^{(c)}$. Lastly, Figure 10(b) shows the estimated partial effect of `doctorco`. It can be seen that an increasing number of doctor consultations leads to a higher expected number of prescribed medications. In addition, an increase in `doctorco` corresponds to lower values of $\vartheta_2^{(1)}$, i.e. smaller dispersion. This is because the parameter $\vartheta_2^{(1)}$ models the dispersion of the discrete response. The effect of `doctorco` on the dependence parameter $\vartheta^{(c)}$ depicted on Figure 10(b) shows that four to six doctor visits correspond to a negative dependence with increasing magnitude, with the trend reversing for six to nine consultations.

Bivariate time-to-event with general censoring scheme

Time-to-event responses may be generally understood as an interval response $Y_{ji} = [L_{ji};\ R_{ji}]$, $j = 1, 2$, where L_{ji} is the left bound and R_{ji} is the right bound of the interval corresponding to the j-th margin. This notation allows us to incorporate a wide range of censoring types. For instance, the case of right-censoring corresponds to $R_{ji} = \infty$. Left-censoring sets the left bound to zero, i.e. $L_{ji} = 0$. For interval-censored observations, $0 < L_{ji} < R_{ji} < \infty$ holds and in case of uncensored observations, we get $L_{ji} = R_{ji}$. The interval responses are accompanied by the indicators δ_U, δ_R , δ_L and δ_I, which indicate whether the observation was uncensored, right-, left- or interval-censored.

The marginal distributions for time-to-event responses implemented in Appendix E (Weibull, log-logistic & log-normal) have been extended to account for the aforementioned general censoring scheme. We illustrate below the log-likelihood of

Table 3: Censoring rates (%) used in the small simulation simulation study shown in Figure 11, $n = 1000$.

	Uncensored	Right	Left	Interval
Margin 1	49.9	19.1	12.4	18.6
(Log-logistic)	9.9	34.4	22.2	33.5
	0.0	38.2	24.7	37.1
Margin 2	49.8	9.9	12.9	27.4
(Log-normal)	10.0	17.6	23.2	49.2
	0.0	19.7	25.7	54.6

the i-th observation of a univariate response:

$$\ell_i = \delta_{U\bullet i} \ln\left[f_\bullet(y_{\bullet i};\ \boldsymbol{\vartheta}_i^{(\bullet)})\right] + \delta_{R\bullet i} \ln\left[S_\bullet(r_{\bullet i};\ \boldsymbol{\vartheta}_i^{(\bullet)})\right] + \\ \delta_{L\bullet i} \ln\left[F_\bullet(l_{\bullet i};\ \boldsymbol{\vartheta}_i^{(\bullet)})\right] + \delta_{I\bullet i} \ln\left[S_\bullet(l_{\bullet i};\ \boldsymbol{\vartheta}_i^{(\bullet)}) - S_\bullet(r_{\bullet i};\ \boldsymbol{\vartheta}_i^{(\bullet)})\right], \quad \bullet \in \{1, 2\}.$$

For bivariate responses, the joint survival function is constructed using a copula as shown in Equation (3), and its corresponding log-likelihood consists of sixteen cases. We conducted a small-scale simulation study using 50 replications with bivariate time-to-event data generated from a Clayton copula. See Table 3 for the censoring rates of each margin used in the simulation study. The synthetic data consisted of $n_{\texttt{train}} = 1000$ and $n_{\texttt{mstop}} = 1000$ observations, which makes the scenario with $p = 1000$ covariates a high-dimensional setting, since a total of $K \times p = 5 \times 1000$ coefficients are to be estimated from the data. The informative covariates are x_1, x_2 and x_4 across the five parameters of the joint survival function.

Results of simulation study

Figure 11 shows the estimated coefficients across a growing number of candidate covariates in the data. Note that only three covariates are informative out of $p = 50$ and $p = 1000$ regressors included in the model and there is a considerable overlap in the parameters where these covariates have an effect. The estimated coefficients indicate that the performance of the boosting algorithm is satisfactory given the amount of uncensored observations in the data (blue boxplots). The degree of shrinkage observed in the dependence parameter $\vartheta^{(c)}$ with $p = 50$ covariates is similar to that observed in the simulations from Appendix E with a 50% of the sample being uncensored. The shrinkage effect increases with a

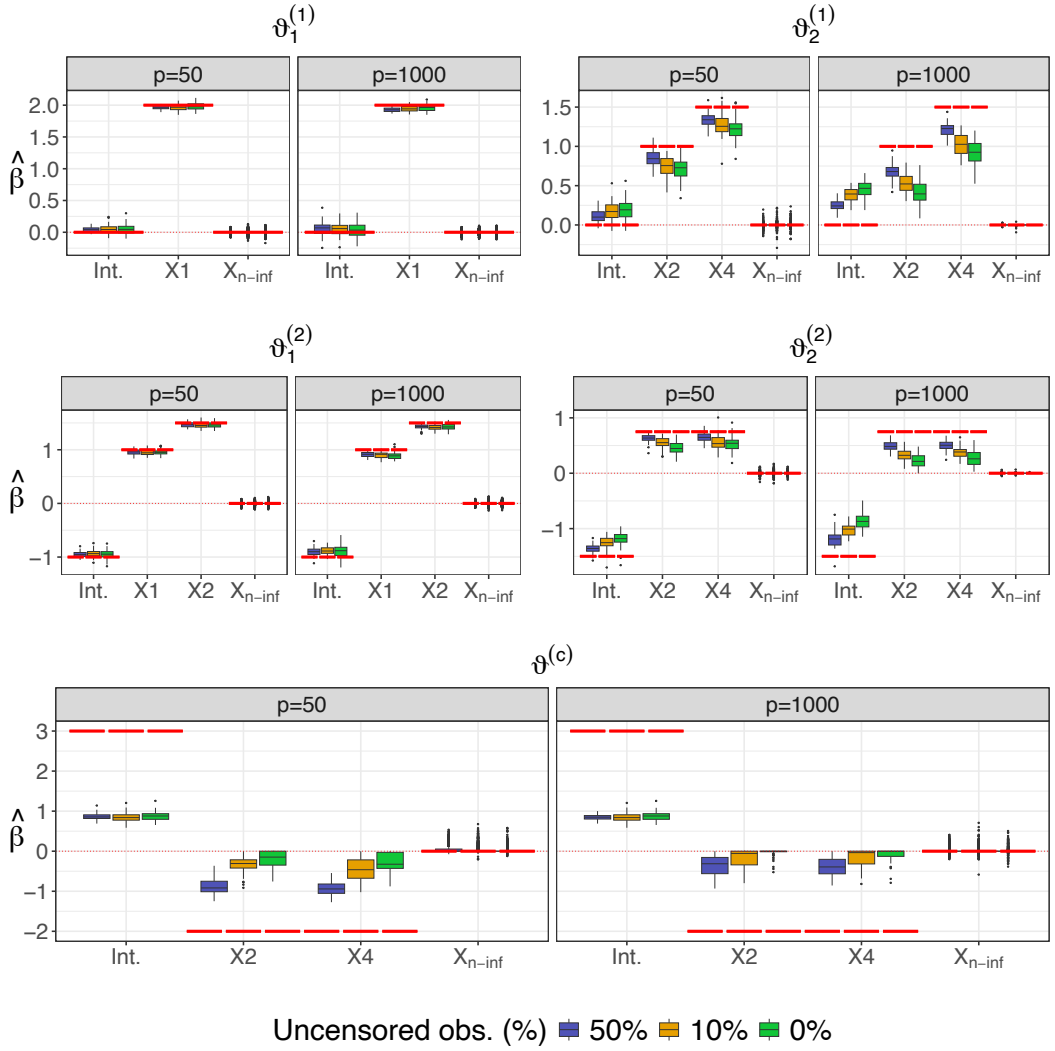

Figure 11: Estimated coefficients of a bivariate time-to-event response with general censoring scheme under different number of covariates in the data ($p = 50$ and $p = 1000$) and fraction of uncensored observations in the sample, $n = 1000$.

decreasing share of uncensored observations. In high-dimensional settings, the algorithm is able to detect the informative covariates in the data across all parameters of the joint survival function with a decreasing share of uncensored observations.

Overall, the results of this preliminary simulation study show promising performance of the boosting algorithm for bivariate time-to-event responses with a general censoring scheme. Future developments will involve the implementation of more copula functions as well as an extensive simulation study with more types of covariate effects as well as different censoring rates. The AREDS data related to the onset of age-related macular degeneration (The AREDS Research

Group, 1999; AMD) is an attractive dataset for a potential application of the proposed method. The proposed boosting algorithm for this type of responses would be the only software routine that combines data-driven variable selection with distributional regression modelling, a wide range of covariate effects, copula functions and marginal distributions. Other approaches from Petti et al. (2022) and Sun and Ding (2019) are able to model bivariate responses with general censoring schemes but do not scale to high-dimensional covariates and cannot conduct data-driven variable selection. A small illustration on how to use the implemented routine of the Clayton copula for responses with general censoring schemes is provided in Appendix E5.

References

Belitz, C., Brezger, A., Klein, N., Kneib, T., Lang, S., and Umlauf, N. (2024). *BayesX - Software for Bayesian inference in structured additive regression models. Version 3.0.2*.
URL http://www.bayesx.org

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, *18*(3-4), 299–321.
URL https://doi.org/10.1177/1471082X17748083

Brezger, A., and Lang, S. (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics & Data Analysis*, *50*(4), 967–991.
URL https://doi.org/10.1016/j.csda.2004.10.011

Briseño Sanchez, G., and Groll, A. (2020). Modelling the effect of rural electrification on employment via component-wise boosted causal distributional regression. In I. I. Garbizu, D.-J. Lee, J. M. Minaya, and M. X. R. Álvarez (Eds.) *Proceedings of the 35th International Workshop Statistical Modelling July 20-24, 2020 Bilbao, Basque Country, Spain*, (pp. 25–30). Universidad del País Vasco / Euskal Herriko Unibertsitatea, Servicio Editorial Argitalpen Zerbitzua.
URL https://dialnet.unirioja.es/servlet/libro?codigo=976089

Briseño Sanchez, G., and Groll, A. (2022). Bivariate mixed binary-survival additive regression modelling. In N. Torelli, R. Bellio, and V. Muggeo (Eds.) *Proceedings of the 36th International Workshop Statistical Modelling July 18-22, 2022 - Trieste, Italy*, (pp. 97–102). EUT Edizioni Università di Trieste.
URL http://hdl.handle.net/10077/33741

Briseño Sanchez, G., Hohberg, M., Groll, A., and Kneib, T. (2020). Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *183*(4), 1553–1574.
URL https://doi.org/10.1111/rssa.12598

Bühlmann, P., and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, *22*(4), 477–505.
URL https://doi.org/10.1214/07-STS242

Chowdhury, M. Z. I., and Turin, T. C. (2020). Variable selection strategies and its

importance in clinical prediction modelling. *Family Medicine and Community Health*, *8*(1), e000262.
URL https://doi.org/10.1136/fmch-2019-000262

Czado, C., and Van Keilegom, I. (2022). Dependent censoring based on parametric copulas. *Biometrika*, *110*(3), 721–738.
URL: https://doi.org/10.1093/biomet/asac067.

Demographic and Health Survey (2023).
URL https://dhsprogram.com/Data/

Duffy, D., Martin, N., and Mathews, J. (1990). Appendectomy in australian twins. *American journal of human genetics*, *47*(3), 590—592.
URL: https://europepmc.org/articles/PMC1683858.
URL https://europepmc.org/articles/PMC1683858

Eilers, P. H. C., and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, *11*(2), 89–121.
URL https://doi.org/10.1214/ss/1038425655

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. D. (2021). *Regression: Models, Methods and Applications*. Springer Berlin Heidelberg.
URL http://dx.doi.org/10.1007/978-3-662-63882-8

Filippou, P., Marra, G., Radice, R., and Zimmer, D. (2022). Estimating the impact of medical care usage on work absenteeism by a trivariate probit model with two binary endogenous explanatory variables. *AStA Advances in Statistical Analysis*, *107*(4), 713–731.
URL http://dx.doi.org/10.1007/s10182-022-00464-6

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232.
URL https://doi.org/10.1214/aos/1013203451

Gioia, V., Fasiolo, M., Browell, J., and Bellio, R. (2022). Additive covariance matrix models: Modelling regional electricity net-demand in great britain.
URL https://arxiv.org/abs/2211.07451

Groll, A., Kneib, T., Mayr, A., and Schauberger, G. (2018). On the dependency of soccer scores – A sparse bivariate poisson model for the UEFA european football championship 2016. *Journal of Quantitative Analysis in Sports*, *14*(2),

65–79.
URL https://doi.org/10.1515/jqas-2017-0067

Han, S., and Vytlacil, E. J. (2017). Identification in a generalization of bivariate probit models with dummy endogenous regressors. *Journal of Econometrics*, *199*(1), 63–73.
URL https://doi.org/10.1016/j.jeconom.2017.04.001

Hans, N., Klein, N., Faschingbauer, F., Schneider, M., and Mayr, A. (2023). Boosting distributional copula regression. *Biometrics*, *79*(3), 2298–2310.
URL https://doi.org/10.1111/biom.13765

Hastie, T., and Tibshirani, R. (1986). Generalized Additive Models. *Statistical Science*, *1*(3), 297 – 310.
URL https://doi.org/10.1214/ss/1177013604

Hepp, T., Schmid, M., Gefeller, O., Waldmann, E., and Mayr, A. (2016). Approaches to regularized regression – A comparison between Gradient Boosting and the LASSO. *Methods of Information in Medicine*, *55*(5), 422–430.
URL https://doi.org/10.3414/ME16-01-0033

Hofner, B., Mayr, A., and Schmid, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, *74*(1), 1–31.
URL https://doi.org/10.18637/jss.v074.i01

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, *36*(2), 299–305.
URL https://doi.org/10.2307/2529982

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2022). *mboost: Model-Based Boosting*. R package version 2.9-7.
URL https://CRAN.R-project.org/package=mboost

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2010). Model-based Boosting 2.0. *Journal of Machine Learning Research*, *11*(71), 2109–2113.
URL http://jmlr.org/papers/v11/hothorn10a.html

Hothorn, T., Kneib, T., and Bühlmann, P. (2013). Conditional Transformation Models. *Journal of the Royal Statistical Society Series B: Statistical Method-*

ology, *76*(1), 3–27.
URL https://doi.org/10.1111/rssb.12017

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2021). *An Introduction to Statistical Learning*. Springer US.
URL https://doi.org/10.1007/978-1-0716-1418-1_6

Jenssen, T.-K., Kuo, W. P., Stokke, T., and Hovig, E. (2002). Associations between gene expressions in breast cancer and patient survival. *Human Genetics*, *111*(4-5), 411–420.
URL https://doi.org/10.1007/s00439-002-0804-5

Jobst, D., Möller, A., and Groß, J. (2024). Gradient-boosted generalized linear models for conditional vine copulas.
URL https://arxiv.org/abs/2406.13500

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, *94*(2), 401–419.
URL https://doi.org/10.1016/j.jmva.2004.06.003

Joe, H. (2014). *Dependence modeling with copulas*. CRC press.
URL https://doi.org/10.1201/b17116

Kalbfleisch, J. D., and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. Wiley.
URL http://dx.doi.org/10.1002/9781118032985

Klein, J. P., and Moeschberger, M. L. (2003). *Survival Analysis: Techniques for Censored and Truncated Data*. Springer New York.
URL https://doi.org/10.1007/b97377

Klein, N. (2024). Distributional regression for data analysis. *Annual Review of Statistics and its Application*, *11*, 321–346.
URL: https://doi.org/10.1146/annurev-statistics-040722-053607.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2014). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *64*(4), 569–591.
URL https://doi.org/10.1111/rssc.12090

Klement, E. P., Mesiar, R., and Pap, E. (2002). Invariant copulas. *Kybernetika*,

38(3), 275–285.
URL http://eudml.org/doc/33582

Kneib, T., Silbersdorff, A., and Säfken, B. (2023). Rage against the mean – A review of distributional regression approaches. *Econometrics and Statistics*, *26*, 99–123.
URL https://doi.org/10.1016/j.ecosta.2021.07.006

Kocherlakota, S., and Kocherlakota, K. (2017). *Bivariate Discrete Distributions*. CRC Press.
URL https://doi.org/10.1201/9781315138480

Kock, L., and Klein, N. (2024). Truly multivariate structured additive distributional regression. *Journal of Computational and Graphical Statistics*, (pp. 1–17).
URL http://dx.doi.org/10.1080/10618600.2024.2434181

Koenker, R. (2005). *Quantile Regression*. Econometric Society Monographs. Cambridge University Press.
URL https://doi.org/10.1017/CBO9780511754098

Lai, C. D., and Balakrishnan, N. (2009). *Continuous Bivariate Distributions*. Springer New York.
URL https://doi.org/10.1007/b101765

Laird, N., and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, *76*(374), 231–240.
URL https://doi.org/10.1080/01621459.1981.10477634

Marra, G., Fasiolo, M., Radice, R., and Winkelmann, R. (2023). A flexible copula regression model with bernoulli and tweedie margins for estimating the effect of spending on mental health. *Health Economics*, *32*(6), 1305–1322.
URL https://doi.org/10.1002/hec.4668

Marra, G., and Radice, R. (2011). A flexible instrumental variable approach. *Statistical Modelling*, *11*(6), 581–603.
URL https://doi.org/10.1177/1471082X1001100607

Marra, G., and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, *112*, 99–113.
URL https://doi.org/10.1016/j.csda.2017.03.004

Marra, G., and Radice, R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, *115*(530), 886–895.
URL https://doi.org/10.1080/01621459.2019.1593178

Marra, G., and Radice, R. (2023). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-6.4.
URL https://cran.r-project.org/package=GJRM

Marra, G., and Radice, R. (2024). A causal transformation model for time-to-event data affected by unobserved confounding.
URL https://arxiv.org/abs/2410.15968

Marra, G., Radice, R., and Zimmer, D. (2024). A unifying switching regime regression framework with applications in health economics. *Econometric Reviews*, *43*(1), 52–70.
URL https://doi.org/10.1080/07474938.2023.2255438

Marra, G., Radice, R., and Zimmer, D. M. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *69*(4), 953–971.
URL https://doi.org/10.1111/rssc.12419

Marra, G., and Wyszynski, K. (2016). Semi-parametric copula sample selection models for count responses. *Computational Statistics & Data Analysis*, *104*, 110–129.
URL https://doi.org/10.1016/j.csda.2016.06.003

Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms: From Machine Learning to Statistical Modelling. *Methods of Information in Medicine*, *53*(6), 419–427.
URL https://doi.org/10.3414/ME13-01-0122

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized Additive Models for Location, Scale and Shape for high dimensional data — A flexible approach based on Boosting. *Journal of the Royal Statistical Society Series C: Applied Statistics*, *61*(3), 403–427.
URL https://doi.org/10.1111/j.1467-9876.2011.01033.x

Nagler, T., and Vatter, T. (2022). *gamCopula: Generalized Additive Models*

for bivariate conditional dependence structures and vine copulas. R package version 0.0-7.
URL https://CRAN.R-project.org/package=gamCopula

Nelder, J. A., and Wedderburn, R. W. M. (1972). Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, *135*(3), 370.
URL https://doi.org/10.2307/2344614

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer New York.
URL https://doi.org/10.1007/0-387-28678-0

Newey, W. K., and Powell, J. L. (1987). Asymmetric least squares estimation and testing. *Econometrica*, *55*(4), 819.
URL https://doi.org/10.2307/1911031

Neyman, J. (1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*(4), 465–472.
URL https://doi.org/10.1214/ss/1177012031

Nikoloulopoulos, A. K., and Karlis, D. (2010). Regression in a copula model for bivariate count data. *Journal of Applied Statistics*, *37*(9), 1555–1568.
URL https://doi.org/10.1080/02664760903093591

Petti, D., Eletti, A., Marra, G., and Radice, R. (2022). Copula link-based additive models for bivariate time-to-event outcomes with general censoring scheme. *Computational Statistics & Data Analysis*, *175*, 107550.
URL https://doi.org/10.1016/j.csda.2022.107550

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
URL https://www.R-project.org/

Riebl, H., Wiemann, P. F. V., and Kneib, T. (2022). Liesel: A probabilistic programming framework for developing semi-parametric regression models and custom bayesian inference algorithms.
URL https://arxiv.org/abs/2209.10975

Rigby, R. A., and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape (with discussion). *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(3), 507–554.
URL https://doi.org/10.1111/j.1467-9876.2005.00510.x

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/9780429298547.

Romeo, J. S., Meyer, R., and Gallardo, D. I. (2018). Bayesian bivariate survival analysis using the power variance function copula. *Lifetime Data Analysis*, *24*(2), 355–383.
URL: http://dx.doi.org/10.1007/s10985-017-9396-1.
URL http://dx.doi.org/10.1007/s10985-017-9396-1

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688–701.
URL https://doi.org/10.1037/h0037350

Stasinopoulos, M., Rigby, R., Voudouris, V., Akantziliotou, C., Enea, M., Kiose, D., and Zeileis, A. (2024). *gamlss: Generalized Additive Models for Location Scale and Shape*. R package version 5.4-22.
URL https://CRAN.R-project.org/package=gamlss

Stock, J. (2001). Instrumental variables in statistics and econometrics. In N. J. Smelser, and P. B. Baltes (Eds.) *International Encyclopedia of the Social & Behavioral Sciences*, (pp. 7577–7582). Oxford: Pergamon.
URL https://www.sciencedirect.com/science/article/pii/B0080430767004484

Strömer, A., Klein, N., Staerk, C., Klinkhammer, H., and Mayr, A. (2023). Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*, *42*(11), 1779–1801.
URL https://doi.org/10.1002/sim.9699

Sun, T., and Ding, Y. (2019). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, *22*(2), 315–330.
URL https://doi.org/10.1093/biostatistics/kxz032

Tadesse, M., and Vannucci, M. (2021). *Handbook of Bayesian Variable Selection*. ISSN. CRC Press.
URL https://doi.org/10.1201/9781003089018

Terza, J. V., Basu, A., and Rathouz, P. J. (2008). Two-stage residual inclusion estimation: Addressing endogeneity in health econometric modeling. *Journal*

of Health Economics, *27*(3), 531–543.
URL https://doi.org/10.1016/j.jhealeco.2007.09.009

The AREDS Research Group (1999). The age-related eye disease study (AREDS): Design implications areds report no. 1. *Controlled Clinical Trials*, *20*(6), 573–600.
URL https://doi.org/10.1016/s0197-2456(99)00031-8

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient Boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, *28*(3), 673–687.
URL https://doi.org/10.1007/s11222-017-9754-6

Tibshirani, R. (2018). Regression shrinkage and selection via the LASSO. *Journal of the Royal Statistical Society: Series B (Methodological)*, *58*(1), 267–288.
URL https://doi.org/10.1111/j.2517-6161.1996.tb02080.x

Tutz, G., and Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer International Publishing.
URL https://doi.org/10.1007/978-3-319-28158-2

Umlauf, N., Klein, N., Zeileis, A., Koehler, M., Simon, T., Stadlmann, S., and Volkmann, A. (2024). *bamlss: Bayesian Additive Models for Location, Scale, and Shape (and Beyond)*. R package version 1.2-4.
URL https://CRAN.R-project.org/package=bamlss

van der Wurp, H., Groll, A., Kneib, T., Marra, G., and Radice, R. (2020). Generalised joint regression for count data: a penalty extension for competitive settings. *Statistics and Computing*, *30*(5), 1419–1432.
URL https://doi.org/10.1007/s11222-020-09953-7

Weisberg, S. (2014). *Applied Linear Regression*. Wiley Series in Probability and Statistics. Wiley.
URL https://doi.org/10.1002/0471704091

Wiemann, P. F., Klein, N., and Kneib, T. (2022). Correcting for sample selection bias in Bayesian distributional regression models. *Computational Statistics & Data Analysis*, *168*, 107382.
URL https://doi.org/10.1016/j.csda.2021.107382

Wood, S. (2023). *mgcv: Mixed GAM Computation Vehicle with Automatic*

Smoothness Estimation. R package version 1.9-1.
URL https://CRAN.R-project.org/package=mgcv

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. CRC press.
URL https://doi.org/10.1201/9781315370279

Wooldridge, J. M. (2015). Control function methods in applied econometrics. *Journal of Human Resources*, *50*(2), 420–445.
URL https://www.jstor.org/stable/24735991

Wyszynski, K., and Marra, G. (2018). Sample selection models for count data in R. *Computational Statistics*, *33*(3), 1385–1412.
URL https://doi.org/10.1007/s00180-017-0762-y

Yee, T., and Moler, C. (2024). *VGAM: Vector Generalized Linear and Additive Models*. R package version 1.1-12.
URL https://CRAN.R-project.org/package=VGAM

Appendix A: Flexible instrumental variable distributional regression (with Supplement)

Joint work with Maike Hohberg, Andreas Groll and Thomas Kneib.

Briseño Sanchez, G., Hohberg, M., Groll, A., & Kneib, T. (2020). Flexible instrumental variable distributional regression. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183 (4), 1553–1574.

Published in the *Journal of the Royal Statistical Society Series A: Statistics in Society*. https://doi.org/10.1111/rssa.12598.

ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Statistics in Society
Series A

J. R. Statist. Soc. A (2020)
183, *Part* 4, *pp.* 1553–1574

Flexible instrumental variable distributional regression

Guillermo Briseño Sanchez,

TU Dortmund University, Germany

Maike Hohberg,

University of Göttingen, Germany

Andreas Groll

TU Dortmund University, Germany

and Thomas Kneib

University of Göttingen, Germany

[Received October 2018. Final revision June 2020]

Summary. We tackle two limitations of standard instrumental variable regression in experimental and observational studies: restricted estimation to the conditional mean of the outcome and the assumption of a linear relationship between regressors and outcome. More flexible regression approaches that solve these limitations have already been developed but have not yet been adopted in causality analysis. The paper develops an instrumental variable estimation procedure building on the framework of generalized additive models for location, scale and shape. This enables modelling all distributional parameters of potentially complex response distributions and non-linear relationships between the explanatory variables, instrument and outcome. The approach shows good performance in simulations and is applied to a study that estimates the effect of rural electrification on the employment of females and males in the South African province of KwaZulu-Natal. We find positive marginal effects for the mean for employment of females rates, negative effects for employment of males and a reduced conditional standard deviation for both, indicating homogenization in employment rates due to the electrification programme. Although none of the effects are statistically significant, the application demonstrates the potentials of using generalized additive models for location, scale and shape in instrumental variable regression for both to account for endogeneity and to estimate treatment effects beyond the mean.

Keywords: Causality; Distributional regression; Generalized additive models for location, scale and shape; Instrumental variable; Treatment effects

1. Introduction

In the potential outcomes framework (Neyman, 1990; Rubin, 1974), causal effects of a binary treatment $D \in \{0, 1\}$ on an outcome variable of interest $Y(D)$ are defined as comparisons between the potential outcome under treatment, $Y_1 = Y(1)$, and the potential outcome without treatment, $Y_0 = Y(0)$, for a common set of units. In practice, we shall be able to observe only either Y_1 or Y_0, depending on the treatment status. When the effect of the treatment is restricted to the mean

Address for correspondence: Guillermo Briseño Sanchez, Fakultät Statistik, Technische Universität Dortmund, Vogelspothsveg 87, Dortmund, Nordhein Westfalen 44227, Germany.
E-mail: briseno@statistik.tu-dortmund.de

of the outcome, the causal effect of the treatment can be reduced to one scalar quantity: the average treatment effect (ATE)

$$\text{ATE} = E(Y_1) - E(Y_0).$$

In heterogeneous populations, the ATE is usually extended to depend on a characteristic X of the individuals of interest, leading to the conditional ATE

$$\text{ATE}(x) = E(Y_1|X=x) - E(Y_0|X=x).$$

Later, we shall consider several characteristics but, for this introductory section, we restrict the notation to the univariate case. The problem with the scalar ATE is that it provides only a rather narrow view of the treatment effect.

Policy makers are often concerned with questions that relate to more general distributional aspects of the variable of interest, such as income inequality, and might then prefer an intervention that lowers the variance or the Gini coefficient of an income distribution over an intervention that has the same ATE but does not reduce inequality. Thus, a more general perspective is to consider potential changes in the complete (conditional) distribution of the outcome, i.e. differences between $\mathcal{D}(Y_1|X=x)$ and $\mathcal{D}(Y_0|X=x)$. In this case, there is not one single scalar treatment effect but rather several treatment effects on various aspects of the distribution of the outcome. This has received considerable interest in the context of quantile regression where conditional treatment effects on specific quantiles $\tau \in (0, 1)$ of the distribution of the outcome can be defined as quantile treatment effects with

$$\text{QTE}_\tau(x) = Q_{Y_1|X=x}(\tau) - Q_{Y_0|X=x}(\tau), \tag{1}$$

where $Q_{Y|X=x|}(\cdot)$ refers to the quantile function of the distribution of the treatment and control potential outcome (see Melly and Wüthrich (2017) for a review of quantile treatment effects in the context of IVs). However, if interest is not only on a specific quantile but also on different features of the distribution such as the variance or the Gini coefficient, it is desirable to estimate the whole conditional distribution directly. With quantile regression, this would require estimating numerous quantile effects and often also dealing with the problem of crossing quantiles.

In this paper, we consider a different approach to evaluate the difference between $\mathcal{D}(Y_1|X=x)$ and $\mathcal{D}(Y_0|X=x)$, where a parametric type of distribution is assumed for Y such that, for example, $\mathcal{D}(Y_1|X=x) = \mathcal{D}_{\vartheta_{Y_1}(x)}$ and $\mathcal{D}(Y_0|X=x) = \mathcal{D}_{\vartheta_{Y_0}(x)}$. This assumes the same type of distribution for Y with and without treatment whereas the parameters of the distribution differ by the treatment (and covariates). One can then either evaluate differences in the parameters with and without treatment directly or derive certain other quantities of interest from the parameters.

To illustrate this point in more detail, assume that $Y_1|X=x \sim \mathcal{N}\{\mu_{Y_1}(x), \sigma^2_{Y_1}(x)\}$ and $Y_0|X=x \sim \mathcal{N}\{\mu_{Y_0}(x), \sigma^2_{Y_0}(x)\}$, i.e. we assume that the outcome of interest follows a normal distribution regardless of whether it receives the treatment or not. However, the parameters of the normal distribution change with treatment, leading to $\vartheta_{Y_1}(x) = (\mu_{Y_1}(x), \sigma^2_{Y_1}(x))$ under treatment and $\vartheta_{Y_0}(x) = (\mu_{Y_0}(x), \sigma^2_{Y_0}(x))$ without treatment. The treatment effect on the mean (given characteristics x) is then given by

$$\text{TE}_\mu(x) = \mu_{Y_1}(x) - \mu_{Y_0}(x). \tag{2}$$

By analogy, the treatment effect on the standard deviation (given characteristics x) is

$$\text{TE}_\sigma(x) = \sigma_{Y_1}(x) - \sigma_{Y_0}(x),$$

but we can also easily derive a treatment effect on, for example, the coefficient of variation as $\sigma_{Y_1}(x)/\mu_{Y_1}(x) - \sigma_{Y_0}(x)/\mu_{Y_0}(x)$ or on the quantiles as in equation (1) by evaluating the inverse

cumulative distribution functions of the corresponding normal distributions. As a consequence, our parametric distributional approach does not provide one single treatment effect but rather a variety of treatment effects on various distributional features that can be derived from the parameters of the outcome distribution with and without treatment. This is particularly so when replacing the normal distribution with more general types of distributions as formalized in generalized additive models for location, scale and shape (GAMLSSs) (Rigby and Stasinopoulos, 2005). The GAMLSS class is a highly flexible model class that allows all parameters of a conditional distribution to vary with covariates and for non-linear relationships between covariates and predictors. It explicitly and parsimoniously models the distribution of the outcome making a GAMLSS more flexible than a linear model. The main advantage over non-parametric models is that conditioning on covariates is inherent in the framework and thus straightforward.

When moving from binary to continuous treatments, the basic set-up that has been discussed so far remains the same but one must specifically determine the status of the treatment variable D before treatment ($D = d_0$) and after treatment ($D = d_1$). For non-linear models focusing on distributional features beyond the mean, the treatment effect then usually explicitly depends on both d_0 and d_1, i.e.

$$\mathrm{TE}_{\vartheta}(x, d_0, d_1) = \vartheta(x, d_1) - \vartheta(x, d_0),$$

where $\vartheta(x, d)$ represents some distributional quantity given characteristics $X = x$ and treatment status $D = d$. For the original treatment status d_0, one often considers the empirical mean from a sample or some representative values of interest. The change in treatment can also be determined differently, e.g. by changing by 1 unit corresponding to the notion of marginal effects. These marginal effects can be calculated at means (marginal effects for means (MEMs)) or at other representative values of covariates, or as average marginal effects (AMEs). Both MEMs and AMEs can then be formulated for different quantities of the distribution. For example, MEMs can be written as

$$\begin{aligned}\text{MEM on mean} &= E\{Y_i(\bar{d}_0 + 1) | X = \bar{x}\} - E\{Y_i(\bar{d}_0) | X = \bar{x}\},\\ \text{AME on mean} &= E\{Y_i(d_{0,i} + 1) | X = x_i\} - E\{Y_i(d_{0,i}) | X = x_i\},\end{aligned}$$

with i indexing the individual and $Y_i(d)$ denoting the outcome for individual i given treatment status $D = d$. Instead of a 1-unit change, changes by 1 standard deviation can also be considered for continuous treatments. For a binary treatment, a marginal effect implies a change in the treatment variable from 0 to 1. Thus, if the sample of individuals is representative for the population, AMEs and MEMs for a binary treatment correspond to the estimated ATE or the estimated conditional ATE depending on whether the covariates are fixed at their observed values or at the mean. For continuous treatments the equivalence between marginal effects and the ATE usually does not hold since the treatment often changes from different baseline levels or by different amounts for each individual.

In the case of perfect randomization and compliance, all treatment effects discussed so far could easily be evaluated by including the treatment variable as an additional covariate in a regression analysis. However, even in randomized control trials (RCTs), compliance of the treatment is often not perfectly observed, calling for an instrumental variable (IV) approach. In other settings, where randomization is not possible, the treatment is biased because of self-selection or other sources of endogeneity. Hence, in many quasi-experimental settings, experimental settings with low compliance or when an explanatory or treatment variable is suspected to be endogenous, an IV can still determine a causal effect. An IV is a variable that affects the treatment or an endogenous covariate but not the outcome and therefore provides information on the causal variation in the response of interest induced by the treatment.

Traditional IV estimators are the Wald estimator or the two-stage least squares estimator 2SLS, where, in the first step, the IV is regressed on the treatment variable (and other covariates) via ordinary least squares and the fitted values from this regression replace then the endogenous treatment variable in the regression specification for the variable of main interest. In a more general setting, 2SLS can not only be applied for treatment effects but for any endogenous covariate, i.e. to determine the causal effect of a covariate that is correlated with the error term. Following Imbens and Angrist (1994) and Angrist *et al.* (1996), such an IV analysis recovers only the local ATE of a certain subgroup (the so-called *compliers*, i.e. it enables correction for deviations from perfect randomization but not for deviations from perfect compliance). In contrast, the two-stage residual inclusion estimator 2SRI includes residuals from the first stage of the IV regression instead of the predicted values and in this way enables the ATE to be targeted instead of the local ATE (Basu *et al.*, 2018). The idea of 2SLS is to use only the treatment part that is independent of the unmeasured confounders to explain the outcome, whereas 2SRI splits the unmeasured confounders into a part that is correlated with the treatment and a part that is not (Guo and Small, 2016). Terza *et al.* (2008) reported good performance of 2SRI when the response variable of main interest does not follow a Gaussian distribution and the expectation of the outcome is related to the covariates by means of a non-linear function. In a way, 2SRI is a form of the control function approach (see Wooldridge (2015) for a review) and was applied to generalized additive models (GAMs) by Marra and Radice (2011). The 2SRI estimator has recently become popular within the field of survival analysis, since the Cox model's hazard rate is connected via a non-linear function to the predictor, making it a directly comparable case with Terza *et al.* (2008). It is also in this strand of literature, where the asymptotic theory for 2SRI has been developed (Jiang *et al.*, 2018; Ying *et al.*, 2019).

We draw on the literature on 2SRI and place it within the GAMLSS framework, not only to estimate treatment effects on the conditional mean of the outcome, but on the whole conditional outcome distribution. In this way, we extend the scope of IV regression towards applications dealing with distributional questions that can be consistently answered by using *one* model. To achieve this goal, we propose an IV estimation procedure within the GAMLSS framework, which we call 2SGAMLSS. The purpose of this paper is twofold: we first analyse the performance of our estimator and, second, we demonstrate what additional insights the GAMLSS framework offers when applying 2SGAMLSS to IV regression.

In a simulation study, we assess the ability of 2GAMLSS to estimate the coefficients of the endogenous variable, the MEMs and AMEs on the mean, and the MEMs and AMEs on the standard deviation of which the second two are not captured by previous approaches. We find that our estimator performs particularly well in all non-linear settings as well as in linear settings where the explanatory and endogenous variables are continuous.

We apply our method to a study on electrification in the South African province of KwaZulu-Natal by Dinkelman (2011) that is presented as a motivating example in Section 2. The *ex post* effect of large infrastructural projects such as electrification can often only be estimated by using IVs, making it a relevant example for the method proposed. The study by Dinkelman (2011) analyses the causal effect of rural electrification on employment rates by using the land gradient as the IV to account for the effect that entering the electrification programme was not at random. Using 2SGAMLSS, we account for non-normal outcomes and non-linearities between treatment and instrument, as well as for the neighbourhood structure between administrative units, and we evaluate the effect of electrification on the whole conditional distribution of employment.

We find that the allocation of an electrification project leads to positive marginal effects on the mean (AMEs and MEMs) for employment rates of females, negative effects for employment

of males, and a reduced conditional standard deviation for both, indicating a homogenization in employment rates. However, these effects are not statistically significant.

The remainder of this paper is structured as follows: Section 2 presents the electrification study and data used, whereas Section 3 briefly reviews existing non-linear IV approaches and introduces 2SGAMLSS. Section 4 performs an extensive simulation study whereas the application on rural electrification is presented in Section 5. Finally, Section 6 concludes.

2. Motivating example

The importance of access to electricity for everyone has gained considerable attention from the international community and is highlighted in the sustainable development goals that were set by the United Nations General Assembly for 2030. Assessing the direct effects of access to electricity on both the individual and the aggregate level is crucial to design electricity programmes to improve livelihoods. Studies show that access to electricity provides positive effects on labour productivity (Lipscomb *et al.*, 2013), household consumption (van de Walle *et al.*, 2017) and individual access to jobs (Grogan and Sadanand, 2013), among others. Given the nature of electricity installations being related to natural settings and political decision making, deriving a causal estimate is difficult. Many studies rely on IV techniques to disentangle such an effect, making electrification an ideal topic for our proposed method.

To apply 2SGAMLSS and to demonstrate what additional information we can draw from it, we rely on rural electrification data from South Africa and replicate a study by Dinkelman (2011) by using the proposed 2SGAMLSS. Using an IV strategy, Dinkelman (2011) estimated the effect of electrification on employment rates for females and males in rural KwaZulu-Natal communities. After Apartheid, during which many households were denied access to electricity, South Africa's electricity utility (Eskom) committed to supplying access to electrification for everyone from 1995 onwards. The following electrification roll-out is considered to suffer from selection bias since flourishing or politically important areas were presumably targeted first. Hence, Dinkelman used in her main analyses an IV strategy with land gradient as the instrument for the allocation of electrification. The idea is that the land gradient is related to project allocation to communities since a higher gradient increases the costs of electrification but is unrelated to the labour market outcomes, which she showed in a placebo experiment.

The data set in the original work is a combination of two census surveys: administrative data on the roll-out and geographical data. The data that we use to replicate her IV analysis are aggregated at the community level and were collected in two waves; one in 1996 (which is used as baseline), and the other in 2001. For our analysis we consider two response variables: the difference in employment between 1996 and 2001 for male and female individuals. The responses, which are denoted by Δ_t prop_male_emp and Δ_t prop_female_emp, are created by taking the proportion of males or females employed with respect to the population in 2001 minus the baseline proportion of 1996.

Table 1 displays the variables that were considered in both the original and our analysis. The endogenous variable (treatment variable) Eskom is a binary indicator that is equal to 1 if a community received an electrification project between 1996 and 2001, and 0 otherwise. The endogenous treatment covers 20% of the $N = 1816$ communities in the sample. The IV Gradient indicates the average land inclination of each community in degrees. Fig. 1 summarizes the example's settings. The outcome variables of interest are the differences in employment rates that are suspected to be influenced by the Eskom electrification programme. Participation in the

Table 1. Summary statistics for the baseline covariates in the analysis†

Variable	*Description*	*Mean*	*Standard deviation*
Δ_t prop_female_emp	Difference in proportion of employment of females	−0.00	0.07
Δ_t prop_male_emp	Difference in proportion of employment of males	−0.04	0.09
Eskom	Electrification project allocation	0.20	0.40
Gradient	Mean land gradient or inclination	10.10	4.89
prop_hh_fem	Proportion of female-led households	0.55	0.13
hh_povrate	Poverty rate	0.61	0.19
sexratio	Sex ratio $N_{\text{females}}/N_{\text{males}}$	1.48	0.28
prop_indianwhite	Proportion of Indian or white adults	0.00	0.01
kms_to_road	Distance (km) to road	37.95	24.57
kms_to_town	Distance (km) to town	38.57	18.12
kms_to_grid	Distance (km) from grid	19.06	13.32
hh_density	Household density	22.05	30.48
prop_hs_male	Proportion of men with high school education	0.06	0.05
prop_hs_fem	Proportion of women with high school education	0.07	0.05
d_prop_flush	Difference in toilet access	0.03	0.08
d_prop_water	Difference in water access	0.01	0.26

†Number of communities $N = 1816$. Number of districts $G = 10$.

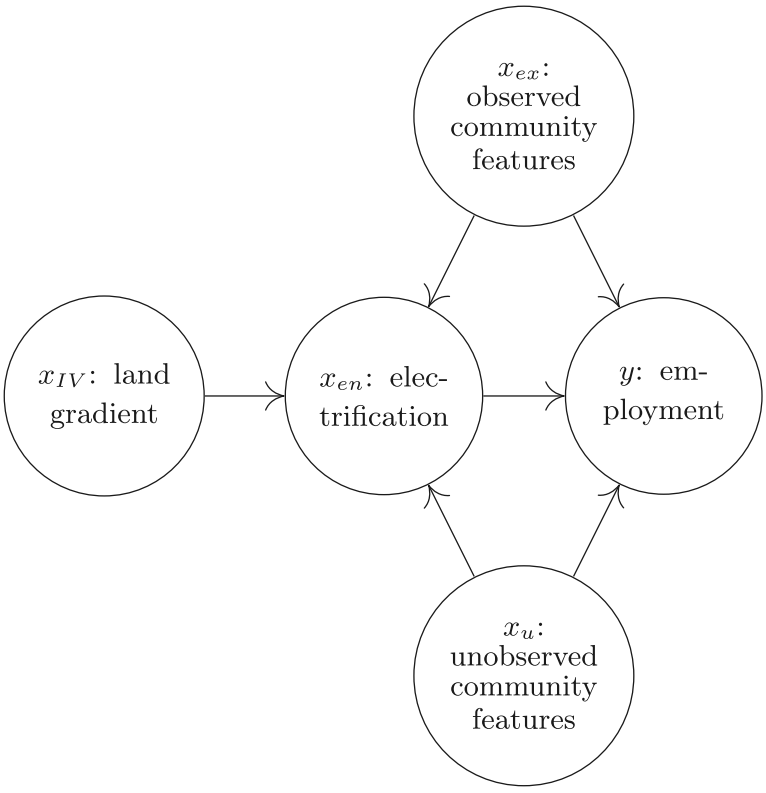

Fig. 1. IV setting for the Eskom programme

programme is prone to endogeneity due to some areas being of higher political interest and is thus instrumented by land gradient.

2SGAMLSS augments the electrification analysis in four ways.

(a) Instead of assuming a linear relationship between the instrument (or other regressors) and the endogenous Eskom treatment, this relationship is modelled flexibly.
(b) In addition to assuming a normal distribution for the employment outcome, we employ a logistic distribution. The *qq*-plots in Fig. B3 in the on-line appendix suggest a slightly better fit of the logistic distribution.
(c) Instead of analysing only treatment effects on the mean, we extend the analysis to causal effects on the standard deviation of each outcome variable.
(d) Instead of modelling the 10 districts in KwaZulu-Natal as fixed effects, we account for the neighbouring structure and employ spatial effects via Gaussian Markov random fields.

3. Methodology

3.1. Non-linear instrumental variable regression

The response variable in our application, the differences in employment rates, is possibly non-normally distributed. When considering non-Gaussian outcomes in the context of generalized linear models, the expectation of the outcome is connected to a linear predictor via a one-to-one response function

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\text{en}}, \mathbf{X}_{\text{ex}}, \mathbf{X}_{\text{u}}) = h(\mathbf{x}_{\text{en}}\beta_{\text{en}} + \mathbf{X}_{\text{ex}}\boldsymbol{\beta}_{\text{ex}} + \mathbf{X}_{\text{u}}\boldsymbol{\beta}_{\text{u}}),$$

where $\mathbf{y}$ is the outcome variable, $\mathbf{X}_{\text{ex}}$ is an $n \times W_{\text{ex}}$ matrix of exogenous variables, $\mathbf{x}_{\text{en}}$ is a column vector and denotes the endogenous treatment variable and $\mathbf{X}_{\text{u}}$ is an $n \times W_{\text{u}}$ matrix of unobservable confounders. In case the model includes several treatment variables, $\mathbf{x}_{\text{en}}$ is replaced by an $n \times W_{\text{en}}$ matrix of endogenous variables $\mathbf{X}_{\text{en}}$. The corresponding unknown regression coefficients are $\boldsymbol{\beta}_{\text{ex}}$ of dimension $W_{\text{ex}} \times 1$, β_{en}, and $\boldsymbol{\beta}_{\text{u}}$ of dimension $W_{\text{u}} \times 1$. The inverse of the response function $h(\cdot)$ is the link function $g(\cdot) = h^{-1}(\cdot)$. For the remainder of this section we assume that $W_{\text{en}} = 1$, i.e. we have only one endogenous variable which is a treatment variable in our example. The reduced form equation of the endogenous explanatory variable can be formulated as

$$\mathbf{x}_{\text{en}} = h_{[1]}(\mathbf{X}_{\text{ex}}\boldsymbol{\delta}_{\text{ex}} + \mathbf{X}_{\text{IV}}\boldsymbol{\delta}_{\text{IV}}) + \boldsymbol{\xi},$$

where $h(\cdot)$ is the conditional expectation of $\mathbf{x}_{\text{en}}$ given the exogenous regressors and IVs, and the subscript '[1]' indicates the first-stage model. The matrix $\mathbf{X}_{\text{IV}}$ is of dimension $n \times W_{\text{IV}}$ and contains the instruments. The exogenous regressors are contained within $\mathbf{X}_{\text{ex}}$. The vectors $\boldsymbol{\delta}_{\text{ex}}$ and $\boldsymbol{\delta}_{\text{IV}}$ are of dimension $W_{\text{ex}} \times 1$ and $W_{\text{IV}} \times 1$ respectively, and they contain the unknown first-stage regression coefficients. The number of elements in $\mathbf{X}_{\text{IV}}$ must be equal to or greater than the number of endogenous regressors and $W_{\text{IV}} \geqslant 1$. The term $\boldsymbol{\xi}$ is a vector of errors of dimension $n \times 1$ that contains information about the unobserved confounders. Replacing the endogenous explanatory variable by its ordinary least squares fitted values no longer isolates the exogenous variation in $\mathbf{x}_{\text{en}}$ from the variable that is generated by $\mathbf{x}_{\text{u}}$. To retrieve the effect of an endogenous explanatory variable on the response in a non-linear context, Terza *et al.* (2008) proposed the following procedure called 2SRI.

(a) Obtain the estimates $\hat{\boldsymbol{\delta}}_{\text{ex}}$ and $\hat{\boldsymbol{\delta}}_{\text{IV}}$ from the first-stage regression by using a generalized linear model algorithm. Define the (pseudo)response residuals as

$$\hat{\boldsymbol{\xi}} = \mathbf{x}_{\text{en}} - h_{[1]}(\mathbf{X}_{\text{ex}}\hat{\boldsymbol{\delta}}_{\text{ex}} + \mathbf{X}_{\text{IV}}\hat{\boldsymbol{\delta}}_{\text{IV}}).$$

(b) For the second-stage model, attach the residuals $\hat{\boldsymbol{\xi}}$ as an additional explanatory variable

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\text{en}}, \mathbf{X}_{\text{ex}}, \hat{\boldsymbol{\xi}}) = h_{[2]}(\mathbf{x}_{\text{en}}\boldsymbol{\beta}_{\text{en}} + \mathbf{X}_{\text{ex}}\boldsymbol{\beta}_{\text{ex}} + \hat{\boldsymbol{\xi}}\beta_{\hat{\boldsymbol{\xi}}})$$

and estimate the unknown coefficients β_{en} and $\boldsymbol{\beta}_{\text{ex}}$, and the coefficient $\beta_{\hat{\boldsymbol{\xi}}}$ via a generalized linear model or other non-linear method.

The estimated residuals $\hat{\boldsymbol{\xi}}$ will then contain information on the unmeasured confounders. However, the regression coefficient $\beta_{\hat{\boldsymbol{\xi}}}$ cannot be employed to explain the effect of the unobserved confounders on the response, since the variation in $\hat{\boldsymbol{\xi}}$ cannot be assigned to any *meaningful* regressor in particular. This is not problematic, since only accounting for $\mathbf{x}_{\text{u}}$'s absence is necessary to obtain a consistent estimate of β_{en}.

The two-stage GAM procedure 2SGAM that was proposed by Marra and Radice (2011) uses the same approach but relaxes the assumption of strictly linear covariate effects in the first and second stage and relates the dependent variable in both stages to an additive predictor (details on the additive predictor are given in the next subsection). The response residuals from the first stage $\hat{\boldsymbol{\xi}}$ enter the second stage as an additional continuous explanatory variable modelled via smooth functions $f_{\hat{\boldsymbol{\xi}}}$ such that

$$\mathbb{E}(\mathbf{y}|\mathbf{x}_{\text{en}}, \mathbf{X}_{\text{ex}}, \hat{\boldsymbol{\xi}}) = h_{[2]}\left\{\mathbf{X}^*_{\text{ex}}\boldsymbol{\beta}^*_{\text{ex}} + \sum_{l=1}^{L} f_l(\mathbf{x}^+_l) + f_{\hat{\boldsymbol{\xi}}}(\hat{\boldsymbol{\xi}})\right\},$$

where the column vectors of $\mathbf{X}^+ = (\mathbf{X}^+_{\text{ex}}, \mathbf{x}^+_{\text{en}})$ and the residuals are modelled as smooth functions and $\mathbf{X}^*_{\text{ex}}$ as linear effects. The model is estimated by using any GAM method, e.g. via `mgcv` in R (Wood, 2017). The smooth estimates of the first-stage residuals account for the influence of the unmeasured confounders; hence we can consistently estimate the effect of the endogenous explanatory variable. Extending this framework in the presence of multiple endogenous regressors results in a total of $W_{\text{en}} > 1$ first-stage regressions. This produces W_{en} vectors of residuals $\hat{\boldsymbol{\xi}}$ that must be modelled either as linear effects via the regression coefficients $\boldsymbol{\beta}_{\hat{\boldsymbol{\xi}}}$ by using 2SRI or as smooth functions by using 2SGAM.

3.2. Generalized additive models for location, scale and shape

Since the response variable follows a certain distribution, we can move away from considering mere mean effects and shift our interest onto the effect on the whole conditional distribution. The GAMLSS method assumes that the observed y_i are conditionally independent and that their distribution can be described by a parametric density $p(y_i|\vartheta_{i1}, \ldots, \vartheta_{iK})$, where $\vartheta_{i1}, \ldots, \vartheta_{iK}$ are K different parameters of the distribution. In the GAMLSS framework, we can specify an equation for each of these parameters of the form

$$g_k(\vartheta_{ik}) = \eta_i^{\vartheta_k} = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\mathbf{x}_{1i}) + \ldots + f_{J_k}^{\vartheta_k}(\mathbf{x}_{J_k i}), \tag{3}$$

where the link function g_k ensures compliance with the parameter space and enables modelling a non-linear relationship between the parameter and the predictor η on the right-hand side of equation (3). The predictor $\eta_i^{\vartheta_k}$ has a structured additive form with $\beta_0^{\vartheta_k}$ denoting the overall level of the predictor and functions $f_j^{\vartheta_k}(\mathbf{x}_{ji})$, $j = 1, \ldots, J_k$, can be chosen to model a range of effects of a vector of explanatory variables $\mathbf{x}_{ji}$.

(a) Linear effects are included via linear functions $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = x_{ji}\beta_j^{\vartheta_k}$, where x_{ji} is a scalar and $\beta_j^{\vartheta_k}$ is a regression coefficient.
(b) Non-linear effects for continuous explanatory variables are captured by smooth functions

$f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(x_{ji})$ where x_{ji} is a scalar. One way of doing this is by using penalized splines (Eilers and Marx, 1996).

(c) Spatial information can be included via $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = f_j^{\vartheta_k}(s_i)$, where s_i is some spatial information such as geographical co-ordinates or administrative units.

(d) For clustered data, random or fixed effects $f_j^{\vartheta_k}(\mathbf{x}_{ji}) = \beta_{j,g_i}^{\vartheta_k}$ can be included with g_i denoting the cluster.

The GAMLSS method has the advantage that it estimates the effects on all parameters of a conditional response distribution that can take basically any parametric form and is thus not bounded to the exponential family. Model estimation can be done by maximum likelihood (Rigby and Stasinopoulos, 2005) or Bayesian methods (Klein *et al.*, 2015). Related R packages are `gamlss` (Stasinopoulos *et al.*, 2017), `GJRM` (Marra and Radice, 2019) and `bamlss` (Umlauf *et al.*, 2018).

3.3. Two-stage generalized additive models for location, scale and shape in instrumental variable regression (2SGAMLSS)

We propose the two-stage GAMLSS method 2SGAMLSS: a procedure that in the first stage performs a distributional regression on the reduced form equation of the endogenous variable x_{en}:

$$g_{\text{en}}(\vartheta_{i,\text{en}}) = \eta_i^{\vartheta_{\text{en}}} = \beta_{0,[1]}^{\vartheta_{\text{en}}} + f_{1,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{\text{IV},i}) + f_{2,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{1i}) + \ldots + f_{J,[1]}^{\vartheta_{\text{en}}}(\mathbf{x}_{J,i}), \tag{4}$$

where $\eta^{\vartheta_{\text{en}}}$ is the structured additive predictor of the conditional expectation of x_{en}, and $g(\cdot) = h^{-1}(\cdot)$ is the link function. The subscript '[1]' indicates that the terms that are specified in equation (4) belong to the first-stage model. The structured additive predictor contains an overall level, as well as effects for the instrument and the remaining exogenous regressors $\mathbf{x}_{1i}, \ldots, \mathbf{x}_{Ji}$. For notational convenience, the subscript k is dropped in equation (4), i.e. a structured additive predictor can be specified for each parameter of the endogenous regressor's distribution. After estimating the regression coefficients in the first-stage model, the conditional expectation of the endogenous regressor and the residuals are computed:

$$\hat{\xi}_i = \mathbf{x}_{\text{en},i} - \mathbb{E}(\mathbf{x}_{\text{en},i} | \hat{\vartheta}_{i,\text{en},1}, \ldots, \hat{\vartheta}_{i,\text{en},K}).$$

Subsequently, all K parameters of the response's density $p(y_i | \vartheta_{i,1}, \ldots, \vartheta_{i,K})$ are regressed on the explanatory variables and the residuals:

$$g_k(\vartheta_{i,k}) = \eta_i^{\vartheta_k} = \beta_{0,[2]}^{\vartheta_k} + f_{1,[2]}^{\vartheta_k}(\mathbf{x}_{1i}) + \ldots + f_{J_k,[2]}^{\vartheta_k}(\mathbf{x}_{J_k i}) + f_{\hat{\xi},[2]}^{\vartheta_k}(\hat{\xi}_i). \tag{5}$$

Here the subscript '[2]' indicates that the components of the K distribution parameters belong to the second-stage model. Note that extending this framework to multiple endogenous regressors results in multiple first-stage models, and having all first-stage residuals attached to the structured additive predictors of the K response distribution parameters. Our proposed procedure resembles that of Marra and Radice (2011) but enables greater flexibility and response distributions that are not members of the exponential family, e.g. zero-inflated distributions.

Especially in an IV setting and in treatment effect evaluation in general, interest often lies in heterogeneous effects. Interaction terms and random coefficients accounting for heterogeneity can be easily included in the second stage. In addition, the GAMLSS method has another notion of heterogeneous effects since they are interpreted conditionally on covariates. One can easily derive AMEs, MEMs or marginal effects at representative values not only for the conditional mean but also for all parameters of the response distribution or other distributional quantities, e.g. the coefficient of variation.

3.4. Confidence intervals

Since 2SGAMLSS relies on two-step estimation, a naive calculation yields intervals that do not necessarily cover their claimed nominal probability, i.e. they will be too narrow. This is because the second-stage regression does not take into account the uncertainty from the quantities that are estimated in the first-stage regression. To represent the uncertainty in the estimated coefficients reliably and to avoid poor coverage, an additional correction must be considered.

Predecessors of 2SGAMLSS have employed a bootstrap pointwise confidence interval correction to restore nominal coverage probabilities. The low coverage probabilities are rectified by employing the joint asymptotic distribution of the GAMLSS maximum likelihood estimators (Stasinopoulos and Rigby, 2007):

$$f(\hat{\boldsymbol{\beta}}|\mathbf{y}) \sim \mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Sigma}),$$

where the vector $\hat{\boldsymbol{\beta}}$ contains the estimates of the unknown regression coefficients $\hat{\boldsymbol{\beta}}$, e.g. obtained via the R routine `GJRM::gamlss()`. The algorithm for obtaining confidence intervals is as follows.

(a) Estimate the first-stage model. Draw a total of N_b random vectors from a multivariate Gaussian distribution: $\mathcal{N}(\hat{\boldsymbol{\beta}}_{[1]}, \hat{\boldsymbol{\Sigma}}_{[1]})$. Calculate all N_b vectors of predictions $\hat{\mathbf{x}}^*_{\mathrm{en},1}, \ldots, \hat{\mathbf{x}}^*_{\mathrm{en},N_\mathrm{b}}$, and their respective residuals $\hat{\boldsymbol{\xi}}^*_1, \ldots, \hat{\boldsymbol{\xi}}^*_{N_\mathrm{b}}$.
(b) Fit the second-stage model N_b times by using the original data attaching the rth vector of residuals. Obtain $\hat{\boldsymbol{\beta}}_{[2],r}$ and $\hat{\boldsymbol{\Sigma}}_{[2],r}$. For each $r = 1, \ldots, N_\mathrm{b}$, draw N_d random vectors from a multivariate Gaussian distribution, i.e. $\mathcal{N}(\hat{\boldsymbol{\beta}}_{[2],l}, \hat{\boldsymbol{\Sigma}}_{[2],l})$ for $l = 1, \ldots, N_\mathrm{d}$.
(c) Calculate the $N_\mathrm{b}N_\mathrm{d}$ fitted values, e.g. $\hat{f}(\mathbf{x}_\mathrm{en})$, and compute the pointwise bootstrap percentile intervals.

Using this procedure, the uncertainty in the residuals $\hat{\boldsymbol{\xi}}$ is accounted for in each of the estimated distribution parameters. We employ the following bootstrap replications for our distributional regression approach: $N_\mathrm{b} = N_\mathrm{d} = 100$.

4. Simulation study

4.1. Simulation set-up

We investigate the pointwise precision of our proposed 2SGAMLSS estimation procedure in a setting that resembles our considered application, i.e. we fit all estimators by assuming a logistic response distribution and binary endogenous treatment variable. For a more detailed description of the data-generating process (DGP), as well as eight alternative scenarios (S1–S8) using different distributions for the response as well as continuous endogenous treatment variable, see the on-line appendix.

We generate a binary endogenous treatment variable by using a structured additive predictor that consists of effects from an unobserved confounder x_u and an instrument x_IV:

$$\eta^{\vartheta_\mathrm{en}} = \phi_1 f_d(x_\mathrm{u}) + \phi_2 f_d(x_\mathrm{IV}).$$

The observation index $i = 1, \ldots, n$ is dropped for notational convenience. The parameters ϕ_1 and ϕ_2 are used to control the strength of the instrument, and the severity of the endogeneity. We specify a strong instrument ($|\rho(x_\mathrm{en}, f_d(x_\mathrm{IV}))| > 0.4$) and severe endogeneity ($|\rho(f_d(x_\mathrm{u}), f_d(x_\mathrm{en}))| > 0.5$). The distributional parameter ϑ_en is obtained by using a response function; then we sample the endogenous treatment x_en from a Bernoulli distribution:

$$\vartheta_{\text{en}} = g_{\text{en}}(\eta^{\vartheta_{\text{en}}})^{-1},$$
$$x_{\text{en}} \sim \text{Ber}(\vartheta_{\text{en}}).$$

Afterwards the additive predictors of the response distribution parameters are created by using effects from x_{en}, x_{u} and some exogenous variables x_{ex}, e.g.

$$\eta^{\vartheta_1} = f_d(x_{\text{ex}_1}) + x_{\text{en}}\beta_{\text{en}} + f_d(x_{\text{u}}),$$
$$\eta^{\vartheta_2} = f_d(x_{\text{ex}_2}) + x_{\text{en}}\beta_{\text{en}} + f_d(x_{\text{u}}).$$

The distributional parameters of the response are obtained by applying the appropriate response function to each predictor. Subsequently, a total of n observations of y are sampled from a logistic distribution:

$$\vartheta_k = g_k(\eta^{\vartheta_k})^{-1},$$
$$y \sim \text{logistic}(\vartheta_1, \vartheta_2).$$

The parameter ϑ_1 corresponds to the mean, whereas the scale parameter ϑ_2 corresponds to a transformation of the variance of the response variable. We created two DGPs by using this framework: one in which the $f_d(\cdot)$ were specified to be strictly linear, and another with $f_d(\cdot)$ as non-linear functions. The specifics of these non-linear functions are detailed in the on-line appendix.

The estimated coefficient of the endogenous variable is compared against 2SLS, 2SRI, 2SGAM, a naive GAMLSS (ignores endogeneity) and full GAMLSS (benchmark, includes the unmeasured confounder) estimators. All the non-linear functions were modelled by using penalized splines. The residuals that were obtained in the first stage of 2SGAMLSS are scaled to have unit variance as recommended in Geraci *et al.* (2016). All estimations were performed in R (R Core Team, 2019).

4.2. Target effects for binary and continuous treatments

For the main setting, we report the median of all estimated endogenous coefficients $\hat{\beta}_{\text{en}}$ on the location and scale parameter. For the remaining scenarios in the on-line appendix with non-linear effects, we focus on pointwise precision quantified by using the root-mean-square error in relation to the true effect of $\mathbf{x}_{\text{en}}$:

$$\text{RMSE}\{\hat{f}(\mathbf{x}_{\text{en}})\} = \sqrt{\left[\frac{1}{N}\sum_{i=1}^{N}\{f(\mathbf{x}_{\text{en},i}) - \hat{f}(\mathbf{x}_{\text{en},i})\}^2\right]},$$

where $\hat{f}(\cdot)$ is the estimated non-linear function evaluation of $\mathbf{x}_{i,\text{en}}$. Additionally, the bias that is incurred by each model is calculated by using

$$\text{bias}\{\hat{f}(\mathbf{x}_{\text{en}})\} = \frac{1}{N}\sum_{i=1}^{N}|\hat{f}(\mathbf{x}_{\text{en},i}) - f(\mathbf{x}_{\text{en},i})|.$$

We report the bias, mean, median and interquantile range IQR, as well as the root-mean-squared error of all Monte Carlo replications of each DGP. These metrics were obtained by using 1000 Monte Carlo replications for sample sizes $N = 500, 2000, 4000$. The uncertainty that is related to estimates obtained via 2SGAMLSS is calculated via coverage probabilities of the bootstrap confidence intervals of $\mathbf{x}_{\text{en}}$. We employ 200 independent data sets using a non-linear DGP. The coverage probabilities were evaluated at confidence levels $\alpha = (0.01, 0.05, 0.1)$.

In addition to the metrics on the coefficient of the endogenous variable, the relative bias between the true and estimated MEMs and AMEs, both on the mean and on the standard deviation sd can be considered. For binary treatments, the MEMs and AMEs are given by

$$\begin{aligned}
\text{MEM on mean} &= E\{Y_i(1)|X_{\text{ex}} = \bar{x}_{\text{ex}}\} - E\{Y_i(0)|X_{\text{ex}} = \bar{x}_{\text{ex}}\},\\
\text{MEM on sd} &= \text{SD}\{Y_i(1)|X_{\text{ex}} = \bar{x}_{\text{ex}}\} - \text{SD}\{Y_i(0)|X_{\text{ex}} = \bar{x}_{\text{ex}}\},\\
\text{AME on mean} &= E\{Y_i(1)|X_{\text{ex}} = x_{\text{ex},i}\} - E\{Y_i(0)|X_{\text{ex}} = x_{\text{ex},i}\},\\
\text{AME on sd} &= \text{SD}\{Y_i(1)|X_{\text{ex}} = x_{\text{ex},i}\} - \text{SD}\{Y_i(0)|X_{\text{ex}} = x_{\text{ex},i}\},
\end{aligned}$$

Equivalently, for continuous treatments MEMs and AMEs can be calculated by

$$\begin{aligned}
\text{MEM on mean} &= E\{Y_i(\bar{x}_{\text{en}} + \text{sd})|X_{\text{ex}} = \bar{x}_{\text{ex}}\} - E\{Y_i(\bar{x}_{\text{en}})|X_{\text{ex}} = \bar{x}_{\text{ex}}\},\\
\text{MEM on SD} &= \text{SD}\{Y_i(\bar{x}_{\text{en}} + \text{sd})|X_{\text{ex}} = \bar{x}_{\text{ex}}\} - \text{SD}\{Y_i(\bar{x}_{\text{en}})|X_{\text{ex}} = \bar{x}_{\text{ex}}\},\\
\text{AME on mean} &= E\{Y_i(x_{\text{en},i} + \text{sd})|X_{\text{ex}} = x_{\text{ex},i}\} - E\{Y_i(x_{\text{en},i})|X_{\text{ex}} = x_{\text{ex},i}\},\\
\text{AME on SD} &= \text{SD}\{Y_i(x_{\text{en},i} + \text{sd})|X_{\text{ex}} = x_{\text{ex},i}) - \text{SD}\{Y_i(x_{\text{en},i})|X_{\text{ex}} = x_{\text{ex},i}\}.
\end{aligned}$$

In the simulation study, we calculate both AMEs and MEMs for binary treatments and focus on the MEMs for continuous treatments with a change of 1 standard deviation to meet the range of the treatment. Focusing on the marginal effects at means or other values has the advantage that we can consider different settings and scenarios of the treatment change, i.e., hypothetically, we could assign different amounts of the treatment to certain individuals to obtain potential outcomes. The advantage of the AMEs is that they provide an overall measure of the actual individuals in the sample. However, AMEs are not adequate if individuals with a certain covariate combination have a very different effect compared with another individual with different covariate values. In practice, we thus recommend calculating AMEs or MEMs or both and, if there is a specific covariate combination that the researcher is interested in, marginal effects at representative values should also be reported.

4.3. Results

Table 2 shows the median of the estimated coefficient $\hat{\beta}_{\text{en}}$ on the response distribution parameters ϑ_1 and ϑ_2 across all Monte Carlo replications. Coefficients estimated by using 2SGAMLSS consistently match those of the benchmark model regardless of DGP and sample size, i.e. estimation of β_{en} is not affected by other linear or non-linear functional forms of the additional covariates. The observed deviations between the medians of 2SGAMLSS and benchmark estimates are very small on both the location (ϑ_1) and scale (ϑ_2) parameters. The estimates of β_{en} produced by the naive GAMLSS method either underestimate or overestimate the covariate's effect on both distributional parameters. This issue is not corrected by increasing the sample size. Standard IV estimation via 2SLS exhibits noticeable deviations from $\hat{\beta}_{\text{en}}$ estimated by the benchmark model as well as our proposed estimator for small sample sizes. Other non-linear IV methods considered such as 2SRI also tend to underestimate the coefficient of the endogenous treatment variable given small sample sizes; see the columns dedicated to $N = 500$ in Table 2. Given larger sample sizes, the estimates from 2SLS, 2SRI, 2SGAM and 2SGAMLSS show minimal differences. However, this behaviour is not observed throughout non-Gaussian responses; see for example the section in the on-line appendix Table A4 dedicated to scenario S6, and S7. Overall, coefficients estimated by using 2SGAMLSS repeatedly match the coefficients delivered by the benchmark model.

The GAMLSS framework enables us to recover the effect of the endogenous regressor on all parameters of the response distribution. The coefficients of x_{en}'s effect on the scale parameter ϑ_2

Table 2. Median $\hat{\beta}_{\text{en}}$ on both distribution parameters ϑ_1 and ϑ_2 of a logistic-distributed response across various sample sizes by using 1000 Monte Carlo replications†

Method	*Results for N = 500*		*Results for N = 2000*		*Results for N = 4000*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\hat{\beta}_{\text{en}}$ *on* ϑ_1						
Naive	1.291	0.674	1.236	0.642	1.277	0.657
2SLS	1.079	0.895	0.923	1.098	1.009	1.068
2SRI	1.028	0.816	0.946	1.098	1.007	1.058
2SGAM	1.125	1.032	0.938	0.954	0.965	1.118
2SGAMLSS	1.188	1.044	0.997	0.994	0.963	1.040
Benchmark	1.099	1.069	0.958	0.976	0.996	0.998
Estimated $\hat{\beta}_{\text{en}}$ *on* ϑ_2						
Naive	1.277	1.232	1.270	1.221	1.273	1.219
2SGAMLSS	0.984	1.009	0.946	0.969	0.944	0.967
Benchmark	1.007	1.014	0.999	1.003	1.002	1.001

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

from 2SGAMLSS exhibit similar values compared with the benchmark model across DGPs and sample sizes. Similarly to the estimates on the location parameter, naive GAMLSS estimates of β_{en} on ϑ_2 remain biased across multiple sample sizes. The lower section of Table 2 shows how our proposed estimator matches the benchmark estimate on the scale parameter across the sample sizes considered. Such consistent estimation of β_{en} is observed across different response distributions; see Table A5 in the on-line appendix. 2SLS, 2SRI and 2SGAM are limited by a constant scale parameter assumption; therefore no coefficient can be recovered for the scale parameter, or any other potential parameter of the conditional response distribution.

If the endogenous variable considered is continuous (scenarios S1–S4), 2SGAMLSS delivers precise effect estimates for x_{en} across all response distribution parameters. This behaviour is observed on both linear and non-linear DGPs. Tables A2 and A3 (in the on-line appendix) show that the mean and median biases of 2SGAMLSS are considerably lower than either naive estimation and other non-linear IV methods (2SRI and 2SGAM) in the location parameter for samples containing around $N = 2000$ observations or more. IQR of the bias of x_{en} also exhibits smaller values for our proposed 2SGAMLSS, indicating a narrower distribution of the bias incurred. For 2SGAMLSS, the values of all metrics considered, i.e. the mean and median bias, the bias's IQR, and RMSE, approach the benchmark values as the sample sizes increases. The metrics on ϑ_2 exhibit overall slightly larger values compared with those for the location parameter ϑ_1, but a trend that favours 2SGAMLSS as the number of observations grows is still noticeable. A somewhat larger sample is essential for 2SGAMLSS to yield consistent effect estimates on the scale parameter. Precise estimation of effects on the scale parameter is crucial since it improves the estimation of heterogeneity in Gaussian responses as in scenario S1 or strictly positive responses as in scenario S4, and overdispersion in count responses as in scenario S3.

Fig. 2 depicts the coverage probabilities of the 2SGAMLSS bootstrap confidence intervals. By setting the bootstrap parameters to $N_{\text{b}} = N_{\text{d}} = 100$, the intervals achieve satisfactory cover-

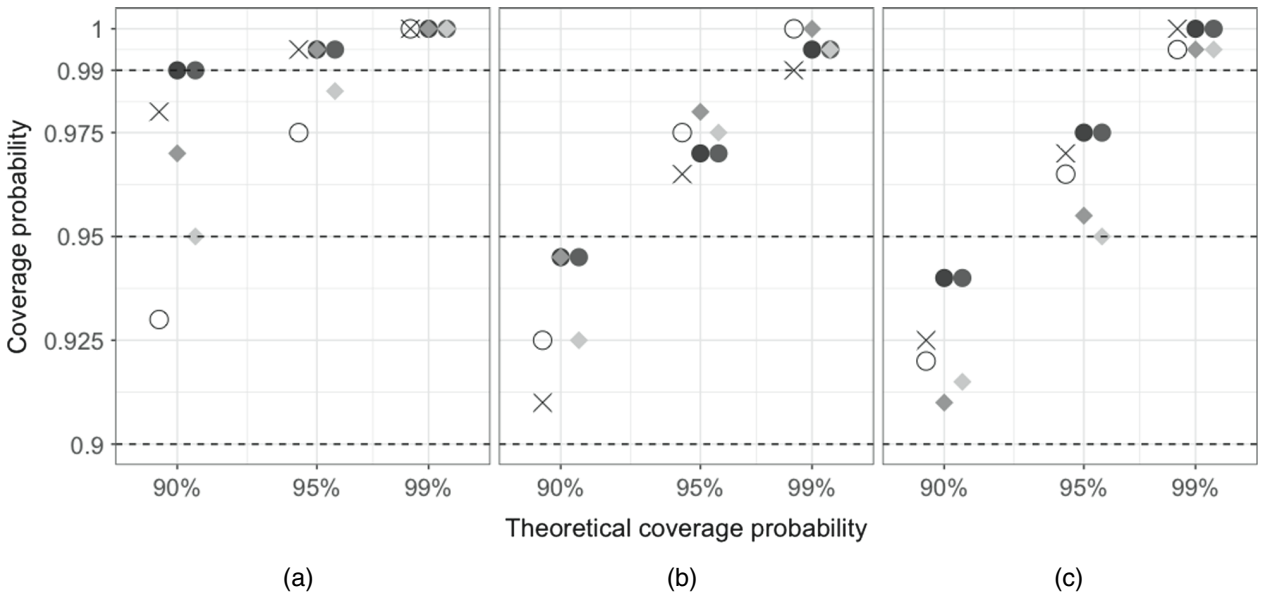

Fig. 2. Coverage probabilities of the bootstrap confidence intervals for β_{en}, and target treatment effects on the mean and standard deviation at various confidence levels (90%, 95%, 99%) across sample sizes (a) $N = 500$, (b) $N = 2000$ and (c) $N = 40000$ by using 200 Monte Carlo replications: ○, location parameter; ●, AME (mean); ●, MEM (mean); ×, scale parameter; ◆, AME (standard deviation); ◆, MEM (standard deviation)

Table 3. Median relative bias of the estimated MEM†

Method	*Results for N = 500*		*Results for N = 2000*		*Results for N = 4000*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\widehat{MEM}$ on the mean						
Naive	−0.291	0.326	−0.236	0.358	−0.277	0.343
2SLS	−0.079	0.105	0.077	−0.098	−0.009	−0.068
2SRI	−0.028	0.184	0.054	−0.098	−0.007	−0.058
2SGAM	−0.125	−0.032	0.062	0.046	0.035	−0.118
2SGAMLSS	−0.180	−0.035	0.003	0.006	0.038	−0.040
Benchmark	−0.099	−0.069	0.042	0.024	0.004	0.002
Estimated $\widehat{MEM}$ on the standard deviation						
Naive	−0.356	−0.414	−0.356	−0.403	−0.358	−0.404
2SGAMLSS	0.013	−0.070	0.033	−0.036	0.029	−0.022
Benchmark	0.004	0.035	0.002	0.041	−0.003	0.043

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

age probabilities of the endogenous treatment effect for the response distribution parameters considered as well as the target treatment effects.

Tables 3 and 4 show the results for the treatment effects on the mean and standard deviation of the outcome, whereas Fig. A2 in the on-line appendix shows boxplots of these. Results for the relative bias in the remaining scenarios are given in the appendix. Neglecting the endogeneity of the treatment variable (naive) leads to considerable bias of both MEMs and AMEs in the mean regardless of the sample size and type of covariate effects (linear or non-linear). In linear

Table 4. Median relative bias of the estimated AME†

Method	*Results for N = 500*		*Results for N = 2000*		*Results N = 4000*	
	Linear	*Non-linear*	*Linear*	*Non-linear*	*Linear*	*Non-linear*
Estimated $\widehat{AME}$ on the mean						
Naive	−0.291	0.326	−0.236	0.358	−0.277	0.343
2SLS	−0.079	0.105	0.077	−0.098	−0.009	−0.068
2SRI	−0.028	0.184	0.054	−0.098	−0.007	−0.058
2SGAM	−0.125	−0.032	0.062	0.046	0.035	−0.118
2SGAMLSS	−0.180	−0.035	0.003	0.006	0.038	−0.040
Benchmark	−0.099	−0.069	0.042	0.024	0.004	0.002
Estimated $\widehat{AME}$ on the standard deviation						
Naive	−0.305	−0.302	−0.301	−0.291	−0.303	−0.293
2SGAMLSS	0.029	−0.012	0.055	0.040	0.053	0.037
Benchmark	−0.000	0.006	0.003	0.004	−0.001	0.003

†The procedures 2SLS, 2SRI and 2SGAM are fitted on a Gaussian distribution, since the logistic distribution cannot be fitted by these procedures, i.e. these estimators are misspecified.

settings, 2SLS outperforms the naive GAMLSS, but in non-linear cases it incurs a sizable bias compared with 2SRI and 2SGAM.

The 2SGAMLSS estimator repeatedly resembles the benchmark estimator in both types of DGP, as well as across the sample sizes considered. As previously mentioned, the GAMLSS framework enables us to derive treatment effects on different distributional quantities of the outcome of interest. The relative bias that is incurred in the AMEs and MEMs on the standard deviation is shown in the lower sections of Table 3 and 4. 2SGAMLSS outperforms the naive estimator at recovering MEMs on the variance in both DGPs and across sample sizes. The naive estimator exhibits the same behaviour as observed in the treatment effects on the mean, i.e. the bias relative to the true value does not benefit from a simpler DGP (linear) or larger sample sizes. As the remaining procedures considered are restricted to estimating the effects on the mean with a constant scale parameter, they are omitted when standard deviation effects are reported.

Although not shown here, we also computed the relative bias for the MEMs and AMEs on the variance of the outcome. The results for the relative bias on the variance qualitatively match those for the standard deviation.

5. Evaluating the effect of rural electrification on employment rates

5.1. Model

To assess how the better performance of 2SGAMLSS in simulation settings translates into a real world scenario, we now come back to the data set on rural electrification in South Africa. We follow the original approach and fit the first-stage distributional regression on the endogenous treatment Eskom by using the instrument Gradient. Regarding the regressors of the first stage, we employ Dinkelman's (2011) most comprehensive specification. This includes community characteristics at the baseline level to control for different growth paths, controls for the different districts and differences in access to water and sanitation. Our approach differs from the

original analysis by modelling the covariates and instrument by using penalized splines instead of strictly linear effects, allowing for data-driven estimation of their (potentially) non-linear functional form. The endogenous covariate Eskom is modelled by using a Bernoulli distribution employing the generalized extreme value link function (by default in `GJRM`) to relate the distribution parameter $\vartheta_1^{\text{Eskom}}$ with the following structured additive predictor:

$$\begin{aligned}\eta_i^{\text{Eskom}} = {} & \beta_{0,[1]}^{\text{Eskom}} + f_{1,[1]}^{\text{Eskom}}(\text{Gradient}_i) + f_{2,[1]}^{\text{Eskom}}(\text{kms\_to\_grid}_i) + f_{3,[1]}^{\text{Eskom}}(\text{kms\_to\_road}_i) \\ & + f_{4,[1]}^{\text{Eskom}}(\text{kms\_to\_town}_i) + f_{5,[1]}^{\text{Eskom}}(\text{hh\_density}_i) + f_{6,[1]}^{\text{Eskom}}(\text{hh\_povrate}_i) \\ & + f_{7,[1]}^{\text{Eskom}}(\text{prop\_hh\_fem}_i) + f_{8,[1]}^{\text{Eskom}}(\text{sexratio}_i) + f_{9,[1]}^{\text{Eskom}}(\text{prop\_indianwhite}_i) \\ & + f_{10,[1]}^{\text{Eskom}}(\text{prop\_hs\_male}_i) + f_{11,[1]}^{\text{Eskom}}(\text{prop\_hs\_fem}_i) + f_{12,[1]}^{\text{Eskom}}(\text{d\_prop\_flush}_i) \\ & + f_{13,[1]}^{\text{Eskom}}(\text{d\_prop\_water}_i) + f_{14,[1]}^{\text{Eskom}}(\text{district}_i).\end{aligned}$$

After estimating the first-stage regression coefficients, we compute the conditional expectation of Eskom and obtain the residuals:

$$\hat{\xi}_i = \text{Eskom}_i - \mathbb{E}(\text{Eskom}_i|\hat{\vartheta}_1^{\text{Eskom}}).$$

The residuals are scaled to have unit variance as in Section 3.1. The quantity $\hat{\xi}$ enters the second-stage predictors as an additional continuous explanatory variable. Our approach further differs from the original study by employing the logistic distribution instead of a Gaussian distribution for the outcomes. For each response separately (i.e. Δ_t prop_female_emp and Δ_t prop_male_emp), we specify a structured additive predictor for the location parameter ϑ_1 with identity link function:

$$\begin{aligned}\eta_i^{\vartheta_1} = {} & \beta_{0,[2]}^{\vartheta_1} + \beta_{1,[2]}^{\vartheta_1}\text{Eskom}_i + f_{2,[2]}^{\vartheta_1}(\hat{\xi}_i) + f_{3,[2]}^{\vartheta_1}(\text{hh\_povrate}_i) + f_{4,[2]}^{\vartheta_1}(\text{hh\_density}_i) \\ & + f_{5,[2]}^{\vartheta_1}(\text{prop\_hh\_fem}_i) + f_{6,[2]}^{\vartheta_1}(\text{prop\_indianwhite}_i) + f_{7,[2]}^{\vartheta_1}(\text{sexratio}_i) \\ & + f_{8,[2]}^{\vartheta_1}(\text{kms\_to\_road}_i) + f_{9,[2]}^{\vartheta_1}(\text{kms\_to\_town}_i) + f_{10,[2]}^{\vartheta_1}(\text{kms\_to\_grid}_i) \\ & + f_{11,[2]}^{\vartheta_1}(\text{prop\_hs\_male}_i) + f_{12,[2]}^{\vartheta_1}(\text{prop\_hs\_fem}_i) + f_{13,[2]}^{\vartheta_1}(\text{d\_prop\_flush}_i) \\ & + f_{14,[2]}^{\vartheta_1}(\text{d\_prop\_water}_i) + f_{15,[2]}^{\vartheta_1}(\text{district}_i).\end{aligned}$$

We specify a structured additive predictor for the scale parameter ϑ_2 with log-link function that features the same set of covariates as the location parameter.

To account better for district heterogeneity, we model the regressor district as a spatial effect by using Markov random fields instead of fixed effects as in the original study. For ϑ_1, the coefficients that are estimated by 2SGAMLSS will reflect the effect of the covariates and residuals on the expectation of the proportion of males/females employment. In the case of the logistic distribution, the conditional variance is a transformation of the scale parameter. The 2SGAMLSS model was fitted using the R package `GJRM` (Marra and Radice, 2019), whereas 2SLS from the original study was fitted using `AER` (Kleiber and Zeileis, 2008).

5.2. First-stage results

In the original study the estimated linear effect of Gradient on Eskom project allocation was negative with $\hat{\beta}_{\text{Gradient}} = -0.0077$. The smooth effect that was estimated in the first stage of 2SGAMLSS that is shown in Fig. 3(a) confirms the existence of an inverse relationship between the instrument and the Eskom project allocation. However, values of Gradient between 0° and 10° land inclination barely have an effect on the structured additive predictor of Eskom. Only when land inclination exceeds 10° will the value of η^{Eskom} start to decrease, *ceteris paribus*. Conventional IV methods such as 2SLS are unable to capture nuances like the range where

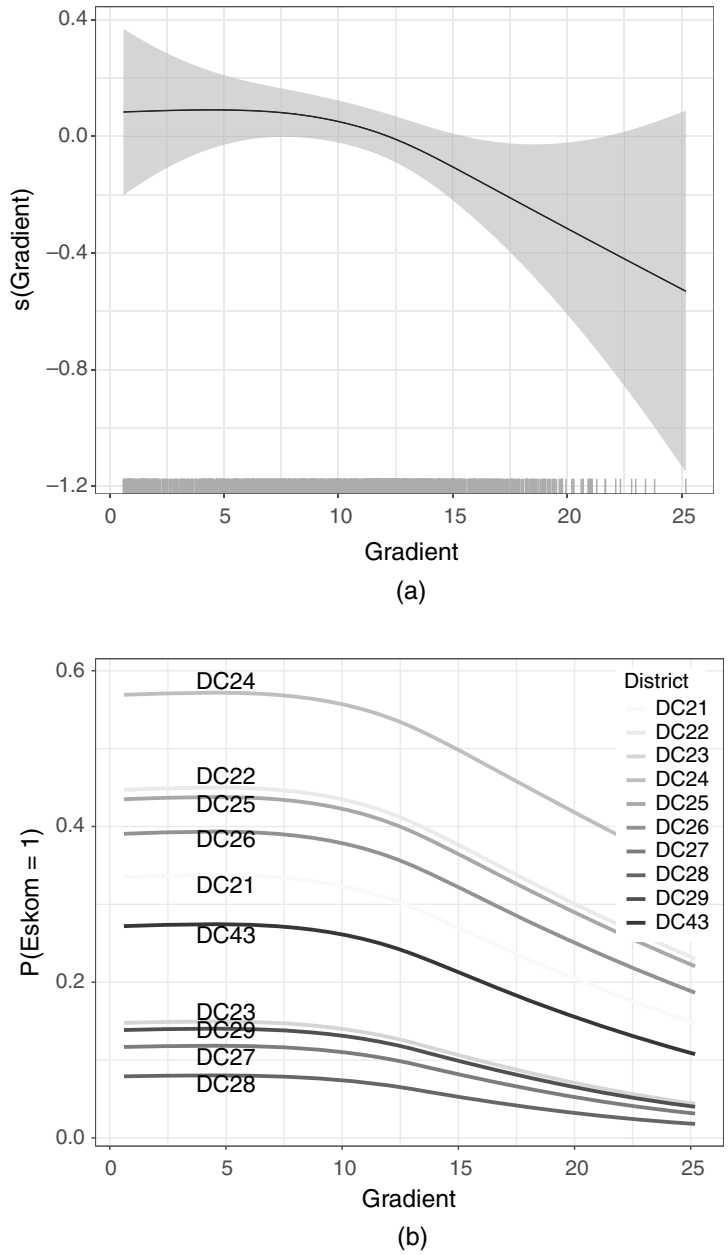

Fig. 3. (a) Estimated smooth effect of the instrument Gradient on the predictor of Eskom with 95% confidence interval and (b) predicted probability of receiving an Eskom project as a function of Gradient across districts DC

Gradient has no effect on the predictor of Eskom. Fig. 3(b) shows the conditional expectation of Eskom as a function of the instrument Gradient, and how the predicted probability of Eskom = 1 varies across the districts of KwaZulu-Natal (see the auxiliary map in Fig. B1 in the on-line appendix to locate neighbouring districts). The inverse relationship between the instrument and Eskom shows steeper descents for some districts (e.g. DC24), but the general trend indicates a decreasing probability of receiving an electrification project as the average land inclination exceeds 10°.

Our proposed approach benefits from the flexible estimation of the instrument's effect, which in turn leads to a better estimate of the first-stage residuals $\hat{\xi}$. The penalized spline estimates of the first-stage residuals on the predictor of ϑ_1 that are shown in Fig. B2 (in the on-line appendix) indicate a strong non-linear effect on the expectation of both response variables. The

Table 5. Regression coefficients (95% confidence intervals), MEMs and AMEs for the electrification data and various IV estimators†

		Results for 2SLS	*Results for 2SGAMLSS*
Δ_t *male employment*			
ϑ_1	Eskom	0.0355	−0.0143
		[−0.0500; 0.2500]	[−0.2615; 0.0156]
	District effects	Fixed	Markov random fields
ϑ_2	Eskom	—	−0.5242
		—	[−1.2302; 0.6032]
	District effects	—	Markov random fields
MEMs or AMEs on mean			−0.0143
			[−0.2615; 0.0156]
MEMs on standard deviation			−0.0010
			[−0.0647; 0.0674]
AMEs on standard deviation			−0.0014
			[−0.2329; 0.0543]
Δ_t *female employment*			
ϑ_1	Eskom	0.0951‡	0.0152
		[0.0500; 0.3000]	[−0.1391; 0.0872]
	District effects	Fixed	Markov random fields
ϑ_2	Eskom	—	−0.8386
		—	[−1.0868; 0.6285]
	District effects	—	Markov random fields
MEMs or AMEs on mean			0.0152
			[−0.1391; 0.0872]
MEMs on standard deviation			−0.0009
			[−0.0523; 0.0499]
AMEs on standard deviation			−0.0012
			[−0.1803, 0.0772]

†All models control for baseline covariates, differences in access to water and sanitation, and district heterogeneity, $N = 1816$. Bootstrap confidence intervals of 2SGAMLSS estimates.
‡$p < 0.1$.

estimates of $\hat{\xi}$ on the structured additive predictor of the scale parameter of both responses also exhibit a non-linear functional form (Fig. B2 in the appendix), which would not have been captured by 2SLS, 2SRI or 2SGAM. It should be noted that these non-linear effects of $\hat{\xi}$ yield no useful interpretation, since the variation in the first-stage residuals cannot be assigned to any explanatory variable. However, the fitted curves validate Dinkelman's (2011) suspicion of the Eskom project's endogeneity.

5.3. Second-stage results

The estimated coefficients for the endogenous treatment in the structured additive predictors of the location and scale parameters of the response distribution are displayed in Table 5. In the original study, the effect of Eskom on both responses was positive and statistically significant for the employment of females. After accounting for possible non-linearities in the covariates, as well as for spatial district heterogeneity, 2SGAMLSS estimates a positive effect of the electricity project allocation on the expectation of the proportion of employment of females. For example, the allocation of an Eskom project will lead to an increase in Δ_t prop_female_emp of 1.52% on average, given that the remaining covariates are held constant. Due to using the identity link and the fact that the location parameter of the logistic distribution equals the conditional mean,

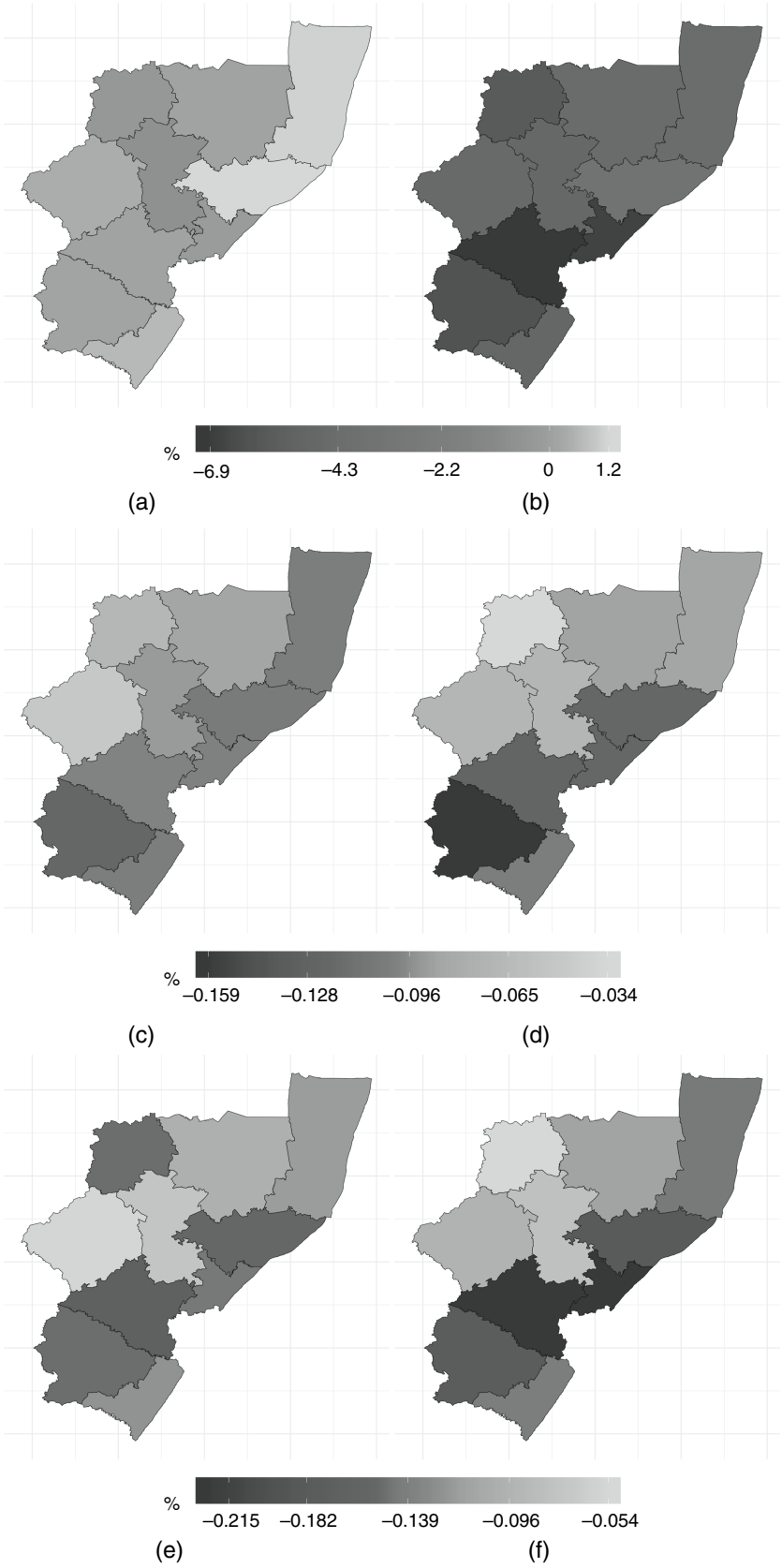

Fig. 4. (a), (b) Estimated MEMs on the mean across districts which correspond to the AMEs for the distribution considered, (c), (d) estimated MEMs on the standard deviation across districts and (e), (f) AMEs on the standard deviation across districts (Δ_t prop_gender_emp$\sim$ logistic(ϑ_1, ϑ_2)): (a), (c), (e) females; (b), (d), (f) males

the effect on ϑ_1 equals the MEMs and AMEs for the mean. The bootstrap confidence intervals indicate that the estimated effect is not significant at the 5% level.

The coefficients for ϑ_2 that were obtained from 2SGAMLSS (logistic) that are shown in Table 5 indicate that the Eskom project allocation has a multiplicative effect on the scale parameter of the response for females of size $\exp(-0.8386) = 0.4323$, given that all other covariates are held constant. Consequently, the MEMs and AMEs for the standard deviation are negative. Note that we decided for the application case to report the effects on the standard deviation due to the scale of the response variable. Variance effects would have been numerically quite small. This means that the conditional standard deviation of the proportion of employment of females for communities that have received an Eskom project is reduced, compared with communities without the electrification project. The coefficient that was estimated by using 2SGAMLSS for Δ_t prop_male_emp suggests that the Eskom project allocation reduces the expected difference in the employment of males by approximately 1.4% and has a multiplicative effect of $\exp(-0.5242) = 0.5920$ on the scale parameter with negative effect on the MEMs and AMEs on the standard deviation.

For a policy maker the combined picture of mean and standard deviation (or variance) effects is of interest. For example, a positive mean effect together with an increase in the standard deviation would have meant that the positive mean effect came mainly through larger benefits for communities that already had higher rates of employment before the programme. Regarding the application, the positive mean and negative standard deviation effect means that a larger increase in employment was experienced by communities that had lower rates of employment before the programme conditionally on covariates. For both men and women, the reduced standard deviation means that the programme led to a homogenization of employment rates between treatment communities compared with control communities, conditionally on covariates. However, the negative mean effect for the employment of males indicates that communities that had higher rates of employment for males before the programme approached the mean rates of employment by experiencing a reduction in rates. Yet, none of the estimated effects for Eskom are significant. The difference between the 2SGAMLSS and the 2SLS estimates for Eskom on the mean of the outcomes could originate from the fact that 2SGAMLSS is based on residual inclusion which tries to recover the ATE, whereas 2SLS estimates the local ATE. Other sources of discrepancy between the fits are 2SGAMLSS's ability to account for possible non-linearities in the covariates' functional form, and the Markov random field representation of the districts' spatial effect.

Fig. 4 depicts the treatment effects for various distributional quantities of both outcomes across the districts of KwaZulu-Natal. The maps for the MEMs and AMEs for the mean indicate that the treatment effect induces a reduction in employment rates for men across all districts. For the employment rates for females, an increase is observed for northern districts and a reduction for a central district. Figs 4(c) and 4(d) display the estimated MEMs for the standard deviation of both outcomes. For women, the MEMs and AMEs for the standard deviation imply a higher degree of homogenization in the east and south than for western districts. For the response for males the estimated treatment effects indicate a reduction in rates of employment accompanied by homogenization of these rates that occurred mostly in the southern districts of KwaZulu-Natal.

6. Concluding remarks

This work proposes an alternative IV estimator which can account for non-normal outcomes, non-linearities between the endogenous variable, instrument and outcome, and can estimate

the treatment effect on the whole conditional distribution and not just the mean. The estimator combines a two-step residual inclusion procedure with the GAMLSS method.

A simulation study shows that, especially in non-linear settings, 2SGAMLSS captures well the coefficient of the endogenous variable. Other non-linear IV methods such as 2SRI and 2SGAM show good performance as well but are restricted to estimation of the mean. For linear settings with Gaussian responses, the results of 2SLS and 2SGAMLSS estimation are very similar. Our IV estimator performs best when both the instrument and the endogenous regressor are continuous. In the presence of endogenous binary variables, the endogenous treatment effect estimated by using 2SGAMLSS repeatedly matches a benchmark estimate for all distribution parameters throughout linear and non-linear settings, regardless of the sample size and the response distribution.

We recommend the implementation of 2SGAMLSS in complex IV settings, where the relationships between outcome, instrument and endogenous regressor(s) are *a priori* unknown. In settings, for which some would claim that interest is solely in the mean, we still suggest using 2SGAMLSS for two reasons. Firstly, once we depart from the Gaussian assumption for the response, there are distributions, such as the Gumbel distribution, whose expected mean depends on more than one parameter. There is no reason why one should be dependent on covariates or one should not. Secondly, more on a philosophical side, we argue that most models should involve considerations beyond the mean to answer research questions from multiple perspectives. We follow Rigby *et al.* (2013) and Kneib (2013) who stated that beyond-the-mean considerations are ubiquitous and models dealing with them should not be regarded as an exception. They gave helpful introductions into beyond-the-mean modelling and mentioned various examples that consider the whole conditional outcome distribution. When estimating the effects of a policy programme, even if the primary interest is in the average effect, any analysis should always be concerned with changes in inequality and whether individuals benefit equally.

We replicated and extended an IV study by Dinkelman (2011) who found positive effects of electrification on employment for both female and male individuals. We found that, in an 'average' community, the effect on employment is positive but only for the employment of females. The effect of electrification on the employment of males was negative. The endogenous treatment variable also impacts the standard deviation of the conditional response distribution, leading in general to a larger reduction in the standard deviation of the employment of males compared with that of females. These statements regarding the treatment effect on the standard deviation of the conditional outcome distributions complement a proper treatment effect evaluation. Effects on the standard deviation are of interest since, first, any treatment produces not just a mean result but varies around that, and second there is no reason to assume that this variation is equal for all people. In addition to the standard deviation, any other distributional feature such as the Gini coefficient or quantiles can be derived from the results, which is essential when we are concerned about inequality and heterogeneity.

These results are of importance not only for extending the GAMLSS applications to IVs but also to policy makers. Infrastructural projects such as electrification are not only the most cost-intensive projects but also those where treatment effects can often only be consistently estimated by using IVs. The method proposed herein enables more exact estimation of the relationships, improving the guidance and justification for policy makers for those projects.

Acknowledgements

The authors thank two referees and the Associate Editor who kindly reviewed earlier versions of the manuscript and provided very valuable suggestions and comments that substantially

improved this work. Maike Hohberg and Thomas Kneib received financial support from the Deutsche Forschungsgemeinschaft within research project KN 922/9-1.

Open access funding enabled and organized by Projekt DEAL.

References

Angrist, J. D., Imbens, G. W. and Rubin, D. B. (1996) Identification of causal effects using instrumental variables. *J. Am. Statist. Ass.*, **91**, 444–455.

Basu, A., Coe, N. B. and Chapman, C. G. (2018) 2SLS versus 2SRI: appropriate methods for rare outcomes and/or rare exposures. *Hlth Econ.*, **27**, 937–955.

Dinkelman, T. (2011) The effects of rural electrification on employment: new evidence from South Africa. *Am. Econ. Rev.*, **101**, 3078–3108.

Eilers, P. H. C. and Marx, B. D. (1996) Flexible smoothing with B-splines and penalties. *Statist. Sci.*, **11**, 89–121.

Geraci, A., Fabbri, D. and Monfardini, C. (2016) Testing exogeneity of multinomial regressors in count data models: does two-stage residual inclusion work? *J. Econmetr. Meth.*, **7**, no. 1.

Grogan, L. and Sadanand, A. (2013) Rural electrification and employment in poor countries: evidence from Nicaragua. *Wrld Devlpmnt*, **43**, 252–265.

Guo, Z. and Small, D. S. (2016) Control function instrumental variable estimation of nonlinear causal effect models. *J. Mach. Learn. Res.*, **17**, 3448–3482.

Imbens, G. W. and Angrist, J. D. (1994) Identification and estimation of local average treatment effects. *Econometrica*, **62**, 467–475.

Jiang, B., Li, J. and Fine, J. (2018) On two-step residual inclusion estimator for instrument variable additive hazards model. *Biostatist. Epidem.*, **2**, 47–60.

Kleiber, C. and Zeileis, A. (2008) *Applied Econometrics with R*. New York: Springer.

Klein, N., Kneib, T., Lang, S. and Sohn, A. (2015) Bayesian structured additive distributional regression with an application to regional income inequality in Germany. *Ann. Appl. Statist.*, **9**, 1024–1052.

Kneib, T. (2013) Beyond mean regression. *Statist. Modllng*, **13**, 275–303.

Lipscomb, M., Mobarak, A. M. and Barham, T. (2013) Development effects of electrification: evidence from the topographic placement of hydropower plants in Brazil. *Am. Econ. J. Appl. Econ.*, **5**, 200–231.

Marra, G. and Radice, R. (2011) A flexible instrumental variable approach. *Statist. Modllng*, **11**, 581–603.

Marra, G. and Radice, R. (2019) GJRM: generalised joint regression modelling. *R Package Version 0.2*. (Available from `https://CRAN.R-project.org/package=GJRM`.)

Melly, B. and Wüthrich, K. (2017) Local quantile treatment effects. In *Handbook of Quantile Regression* (eds R. Koenker, V. Chernozhukov, X. Hu and L. Peng), pp. 145–164. Boca Raton: Chapman and Hall–CRC.

Neyman, J. (1990) On the application of probability theory to agricultural experiments: essay on principles, section 9 (Engl. transl. D. Dabrowska and T. Speed). *Statist. Sci.*, **5**, 465–472.

R Core Team (2019) *R: a Language and Environment for Statistical Computing*. Vienna: R Foundation for Statistical Computing.

Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.

Rigby, R., Stasinopoulos, D. and Voudouris, V. (2013) A comparison of GAMLSS with quantile regression. *Statist. Modllng*, **13**, 335–348.

Rubin, D. B. (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J. Educ. Psychol.*, **66**, 688–701.

Stasinopoulos, M. D. and Rigby, R. A. (2007) Generalized additive models for location scale and shape (GAMLSS) in R. *J. Statist. Softwr.*, **23**, no. 7, 1–46.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V. and De Bastiani, F. (2017) *Flexible Regression and Smoothing: using GAMLSS in R*. Boca Raton: CRC Press.

Terza, J. V., Basu, A. and Rathouz, P. J. (2008) Two-stage residual inclusion estimation: addressing endogeneity in health econometric modeling. *J. Hlth Econ.*, **27**, 531–543.

Umlauf, N., Klein, N. and Zeileis, A. (2018) BAMLSS: Bayesian additive models for location, scale, and shape (and beyond). *J. Computnl Graph. Statist.*, **27**, 612–627.

Venables, W. N. and Ripley, B. D. (2002) *Modern Applied Statistics with S*, 4th edn. New York: Springer.

van de Walle, D., Ravallion, M., Mendiratta, V. and Koolwal, G. (2017) Long-term gains from electrification in rural India. *Wrld Bank Econ. Rev.*, **31**, 385–411.

Wood, S. N. (2017) *Generalized Additive Models: an Introduction with R*, 2nd edn. Boca Raton: Chapman and Hall–CRC.

Wooldridge, J. M. (2015) Control function methods in applied econometrics. *J. Hum. Resour.*, **50**, 420–445.

Ying, A., Xu, R. and Murphy, J. (2019) Two-stage residual inclusion for survival data and competing risks—an instrumental variable approach with application to SEER-Medicare linked data. *Statist. Med.*, **38**, 1775–1801.

Supporting information

Additional 'supporting information' may be found in the on-line version of this article.

A Details on simulation study

A.1 Design

The generation of synthetic data starts from a set of multiple additive predictors. By means of link functions, different regressors are associated to the parameters of arbitrary response distributions. We generate scenarios with the following properties: (1) The relationship between the regressors and response's predictor exhibits non-linear functional forms. (2) The regressor x_{en} is endogenous due to correlation with an omitted variable x_u. (3) An instrument x_{IV} induces *sufficient* variation in x_{en} and has no effect on the response after conditioning on x_{en}. (4) The instrument is independent of the unmeasured confounder x_u. Each scenario is studied using a total of C Monte Carlo Replications. The predictors are generated as follows:

- Sample N observations of the exogenous regressors, unobserved confounder, and the instrument from independent, univariate uniform distributions

$$x_{ex_1} \sim \mathcal{U}[0,1] \qquad x_{ex_2} \sim \mathcal{U}[0,1] \qquad x_u \sim \mathcal{U}[0,1] \qquad x_{IV} \sim \mathcal{U}[0,1]$$

- Generate the additive predictor of the endogenous regressor using:

$$\eta^{\vartheta_{en}} = \phi_1 f_d(x_u) + \phi_2 f_d(x_{IV})$$

  The parameters ϕ_1, ϕ_2 are used to control the strength of the instrument: $\rho(f_d(x_{IV}), x_{en})$, and the severeness of the endogeneity: $\rho(f_d(x_{en}), f_d(x_u))$. The simulation study is performed using a *strong* instrument x_{IV} ($|\rho(x_{en}, f_d(x_{IV}))| > 0.4$) and *severe* endogeneity ($|\rho(f_d(x_u), f_d(x_{en}))| > 0.5$). The endogenous regressor is allowed to have either a continuous or binary form.

- Apply the response function to $\eta^{\vartheta_{en}}$ to obtain ϑ_{en}. Sample N observations of x_{en} from the desired distribution

$$\vartheta_{en} = g_{en}(\eta^{\vartheta_{en}})^{-1}$$
$$x_{en} \sim p(x_{en}|\vartheta_{en})$$

- Using the sampled x_{en}, construct the additive predictors of the response distribution parameters by applying nonlinear functions to the covariates

$$\eta^{\vartheta_1} = f_d(x_{ex_1}) + f_d(x_{en}) + f_d(x_u)$$
$$\eta^{\vartheta_2} = f_d(x_{ex_2}) + f_d(x_{en}) + f_d(x_u)$$

- Obtain the response distribution parameters by applying the response function to each predictor. Sample N observations of y from the desired conditional response distribution

$$\vartheta_k = g_k(\eta^{\vartheta_k})^{-1}$$
$$y \sim p(y|\vartheta_1, \ldots, \vartheta_K).$$

The non-linear functions $f_d(\bullet)$ are taken from Table A13 and are shown in Figure A3. Similar to the study conducted in Marra and Radice (2011), we re-scale all functions $f_d(\bullet)$ to lie within $[0, 1]$. We consider an alternative data generating process (DGP) consisting of purely linear effects is considered, i.e. instead of non-linear functions $f_d(\bullet)$, each of the simulated covariates enters the linear predictors of x_{en} and y directly. This formulation is used to investigate the performance of the 2SGAMLSS procedure in cases where the underlying functional forms are linear.

Six different models are estimated on each sample assuming the correct response distribution, e.g. if the response follows a Gaussian distribution, then the GAMLSS models also assume a Gaussian distribution. We fit the standard 2SLS using:

$$\begin{aligned}\mathbf{x}_{en} &= \beta_{0,[1]} + \mathbf{x}_{IV}\boldsymbol{\beta}_{IV,[1]} + \mathbf{x}_{ex}\boldsymbol{\beta}_{ex,[1]},\\ \mathbf{y} &= \beta_{0,[2]} + \mathbf{x}_{ex}\boldsymbol{\beta}_{ex,[2]} + \hat{\mathbf{x}}_{en}\boldsymbol{\beta}_{en,[2]}.\end{aligned}$$

The 2SRI includes the first-stage residuals as shown in Section 3.1. On the other hand, 2SGAM features a similar specification as 2SRI, but accounts for possible non-linearities in the covariates. The 2SGAMLSS is specified using structured additive predictors for the parameters of the endogenous regressor, as well as the response distribution parameters:

$$\begin{aligned}\vartheta^{en} &= \beta_{0,[1]}^{\vartheta_{en}} + f_{1,[1]}^{\vartheta_{en}}(\mathbf{x}_{IV}) + f_{2,[1]}^{\vartheta_{en}}(\mathbf{x}_{ex}),\\ \vartheta_k &= \beta_{0,[2]}^{\vartheta_k} + f_{1,[2]}^{\vartheta_k}(\mathbf{x}_{ex}) + f_{2,[2]}^{\vartheta_k}(\mathbf{x}_{en}) + f_{3,[2]}^{\vartheta_k}(\hat{\boldsymbol{\xi}}).\end{aligned}$$

where the residuals are computed as in Section 3. We include a naive GAMLSS model in which the endogeneity of $\mathbf{x}_{en}$ is ignored:

$$\vartheta_k = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\mathbf{x}_{ex}) + f_2^{\vartheta_k}(\mathbf{x}_{en}).$$

Lastly, we fit a full model in which the unobserved confounder $\mathbf{x}_u$ is included in all additive predictors of the response distribution parameters, i.e. it performs a regression on all of the true components of the response distribution parameters. This model represents our best possible or benchmark estimation of $\mathbf{x}_{en}$'s effect:

$$\vartheta_k = \beta_0^{\vartheta_k} + f_1^{\vartheta_k}(\mathbf{x}_{ex}) + f_2^{\vartheta_k}(\mathbf{x}_{en}) + f_3^{\vartheta_k}(\mathbf{x}_u).$$

The subscripts [1], and [2] denote the first and second stage of the respective IV method. All of the non-linear functions are modelled using P-splines. The procedures 2SLS, 2SRI, 2SGAM and 2SGAMLSS are estimated in their respective two-step fashion. The residuals obtained in the first stage of 2SGAMLSS are scaled to have unit variance as recommended in Geraci et al. (2016).

We focus primarily on the point-wise precision of the proposed method. The median of all estimated coefficients $\hat{\beta}_{en}$ is computed across different sample sizes, DGPs and Monte Carlo replications. In presence of non-linear effects, we focus primarily on the point-wise precision quantified using the Root Mean Square Error (RMSE) in relation to the true effect of x_{en}:

$$\text{RMSE}(\hat{f}(\mathbf{x}_{en})) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(f(\mathbf{x}_{en,i}) - \hat{f}(\mathbf{x}_{en,i}))^2},$$

where $\hat{f}(\bullet)$ is the estimated non-linear function evaluation of $x_{en,i}$. Additionally, the bias incurred by each model is calculated using:

$$\text{Bias}(\hat{f}(\mathbf{x}_{en})) = \frac{1}{N}\sum_{i=1}^{N}|(\hat{f}(\mathbf{x}_{en,i}) - f(\mathbf{x}_{en,i})|.$$

We report the bias' mean, median, and inter-quantile range (IQR), as well as the RMSE of all Monte Carlo replications of each data generating process in these settings (scenarios S1-S4).

For the treatment effects (AME & MEM) on the mean and standard deviation, we compute the median bias relative to the true value over all Monte Carlo replications, e.g.

$$\text{Bias}_{\%}(\widehat{AME}_{\bullet}) = Median\left(\frac{AME_{True} - \widehat{AME}_{\bullet}}{AME_{True}}\right).$$

where the index " $\bullet$ " denotes the respective distributional quantity. We compute both AME and MEM for scenarios with discrete treatments (S5-S9), whereas for scenarios with continuous treatment (S1-S4) we only consider the MEM with treatment at means plus one standard deviation against the treatment at means.

These metrics are obtained using 1000 Monte Carlo replications for the following sample sizes: $N = 500, 2000, 4000$. The uncertainty related to estimates obtained via 2SGAMLSS is calculated via coverage probabilities of the bootstrap confidence intervals of x_{en}. We employ 200 independent data sets and compute the coverage probability at the following confidence levels: $\alpha = (0.01, 0.05, 0.1)$.

The non-linear functions $f_d(\bullet)$ are estimated using P-splines. The functions are approximated using a B-spline basis on $q = 20$ equidistant knots, and cubic polynomials (degree $l = 3$). All GAMLSS models are estimated using the `R` package `GJRM` (Marra and Radice, 2019). The 2SGAM is estimated using `mgcv` (Wood, 2017). For some response distributions, the 2SRI model is estimated using `MASS` (scenarios S3 & S7). Overall a total of nine distributional scenarios (Table A1) are be investigated using the aforementioned sample sizes and DGPs. All calculations are performed in `R` (R Core Team, 2019)

Table A1: Studied distributional scenarios with their respective distribution parameters and link functions.

Scenario	Type	$\mathbf{x}_{en}$	ϑ_{en}	$\mathbf{y}$	ϑ_y	Link
S1	Continuous	Gaussian	μ	Gaussian	μ, σ	Identity, log
S2		Gaussian	μ	Binomial	μ	Logit
S3		Gaussian	μ	Negative Binomial	μ, σ	log, log
S4		Gaussian	μ	Gamma	μ, σ	log, log
S5	Binary	Bernoulli	μ	Gaussian	μ, σ	Identity, log
S6		Bernoulli	μ	Binomial	μ	Logit
S7		Bernoulli	μ	Negative Binomial	μ, σ	log, log
S8		Bernoulli	μ	Gamma	μ, σ	log, log
S9		Bernoulli	μ	Logistic	μ, σ	Identity, log

A.2 Results

In scenarios with continuous endogenous regressors (S1-S4), precision indicators (RMSE and bias) are measured "across-the-function" for the effect of x_{en} on the location and scale parameters (ϑ_1 and ϑ_2). For the scenarios with binary endogeneity (S5-S9), we compare the estimated regression coefficients for the endogenous regressor ($\hat{\beta}_{en}$) from each type of estimation against the benchmark model.

Figure A1 exemplifies the improved precision of our proposed approach. The graphic shows boxplots of the incurred RMSE by each estimation procedure on x_{en}'s effect on the location parameter of a Negative Binomial response distribution (S3) across different sample sizes and DGPs. The RMSE of all models, except for the 2SLS, share a common range when the sample is small ($N = 500$), although a trend that favours the 2SGAMLSS procedure is pronounced. Figure A1 highlights the severity of the RMSE incurred by 2SLS under non-Gaussian responses. IV procedures tailored for non-Gaussian responses such as 2SRI also show higher RMSE values under non-linear DGPs. On the other hand, the RMSE of the 2SGAMLSS closely resembles that of the benchmark model throughout the studied sample sizes. Even under a linear data generating process, the metrics recorded for 2SRI & 2SGAM remain slightly larger than those of 2SGAMLSS. As the sample size increases, the 2SGAMLSS and benchmark models are able to further reduce their RMSE, whereas the remaining estimation procedures only reduce the spread of their RMSE. Once the sample size is relatively large ($N \approx 2000$), estimates obtained from the proposed 2SGAMLSS will resemble those of the the benchmark model. The graphical results from scenario S3 (Figure A1) indicate that the implementation of 2SGAMLSS produces consistent estimates in the presence of non-linear effects and endogeneity. This pattern is observed for all scenarios of distributions for the endogenous regressor and the response variable, albeit to a varying degree.

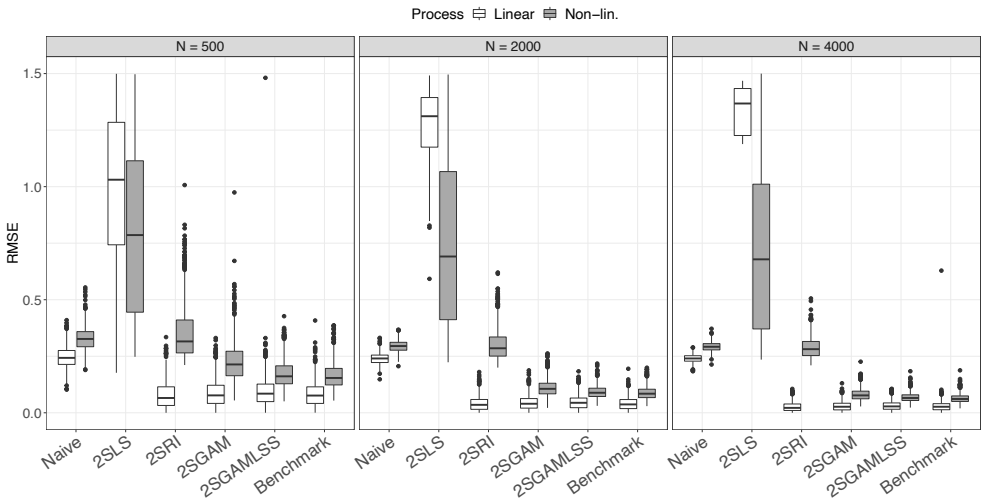

Figure A1: RMSE of x_{en}'s effect on the location parameter ϑ_1 across linear and non-linear DGPs (scenario S3).

A.2.1 Estimation given a continuous treatment

Table A2 shows both mean, median and IQR of the bias, as well as the incurred RMSE of scenarios in which the endogenous explanatory variable is continuous. The results show that under scenario S1 the 2SGAMLSS requires more observations to reduce the bias of x_{en}'s estimated effect on ϑ_1 (both in linear and non-linear DGPs). Scenario S1 is the only case in which a naive GAMLSS produced estimates of similar precision to that of the 2SGAMLSS and the benchmark. This is most likely due to the simple assumption of a Gaussian response. The poor RMSE performance of the 2SLS estimation can be largely attributed to the bias incurred by the procedure. The variance of the estimation is likely to remain small in these situations. Additionally, the employed DGP simulates the predictors of the location and scale parameters as non-linear functions of the explanatory variables. A GAMLSS estimation will be able to accurately assign the variation in x_{en} to each distribution parameter, whereas 2SLS will combine x_{en}'s variation of ϑ_1 with that of ϑ_2, hence the larger bias and RMSE. The bias and RMSE of 2SRI and 2SGAM also remain larger than those of 2SGAMLSS. In circumstances where the scale parameter is unaffected by the explanatory variables and their effects remain strictly linear, 2SLS's precision will be similar to that of the GAMLSS models.

The 2SGAMLSS noticeably outperforms the rest of the estimators in terms of precision (notice an already lower bias and RMSE on $N = 500$). Under over-dispersed count responses (scenario S3), the benefits of 2SGAMLSS become even more evident. Scenario S4 was the only studied scenario (with continuous endogenous covariate) in which the 2SGAMLSS failed to achieve similar precision as the benchmark using small samples. Results similar to scenarios S2 and S3 (given nonlinear DGPs) were seen after taking N towards 2000, and more observations. For the non-Gaussian responses, bias and RMSE of the 2SGAMLSS are dramatically reduced by having more observations in the sample. For example, the section dedicated to scenario S3 in Table A2 supports the necessity of large samples when implementing instrumental variable estimation with distributional regression, especially when the underlying DGP exhibits non-linearities between the covariates and outcome. The sections corresponding to $N = 500$ in Table A2 show how a small sample affects the quality of 2SGAMLSS estimation. RMSE, as well as mean and median bias are often larger than those of the naive estimation. A (relatively) small sample does not allow for enough variation in the first stage which will lead to biased estimates on the second stage. We also found that the distribution of the bias incurred by 2SGAMLSS remained symmetrical in the scenarios depicted in Table A2. When fitting non-Gaussian responses, neglecting the endogeneity in x_{en} results in largely biased and inconsistent estimates whereas the 2SGAMLSS's precision converges towards that of a full model (benchmark). Adding observations mostly contributes to a noticeable improvement of our approach. Increasing the sample size reduces the RMSE for all models eventually reaching a non-zero lower bound for the 2SLS, 2SRI, 2SGAM, and naive estimation. On the other hand, the RMSE of 2SGAMLSS approaches zero when using a very large sample on most of our simulated scenarios where x_{en} was continuous (S1-S3).

Estimation of the effect a continuous x_{en} on ϑ_2 using 2SGAMLSS shows the same traits as its estimation on the location parameter ϑ_1. Table A3 indicates that under Gaussian and non-Gaussian responses, the 2SGAMLSS produces estimates that also resemble those of the benchmark model in the studied DGPs. In scenario S1, the bias on ϑ_2 incurred by 2SLS remains large and constant. This is attributed to 2SLS being a mean-regression procedure, i.e. ϑ_2 is treated as a nuisance parameter and is not modelled as function of the explanatory variables. This limitation also holds for both 2SRI and 2SGAM. A naive GAMLSS estimation is also prone to deliver biased estimates on the scale parameter. Although the more flexible modelling framework would allow for a regression on the scale parameter ϑ_2 (e.g. by including non-linear effects), the endogeneity in x_{en} prevents it from producing consistent estimates. The empirical distribution of the bias using small samples is heavily skewed and prone to outliers (see the difference between mean and median bias), hence comparing the median of all procedures will provide a better picture of their precision. The estimation methods behave similarly as in the location parameter, although the bias and RMSE remain far from zero in cases of non-Gaussian responses.

Results for the target treatment effects also show a better performance from the 2SGAMLSS compared to other methods. Note that in scenarios with a continuous treatment (S1-S4) we are only measuring the bias at one point of the treatment effect, at means plus one standard deviation compared to at the treatment at means, instead of the entirety of the effect curve at representative values, e.g. treatment at means against the entire range of treatment values.

A.2.2 Estimation given a binary treatment

In scenarios with binary endogenous regressors, the 2SGAMLSS estimation outperforms the naive model, especially if the employed sample is large. In contrast to Table A2, a considerably larger sample is essential for 2SGAMLSS to yield consistent estimates of x_{en}'s effect on the scale parameter. Throughout these settings (S5-S9), the estimated treatment effect given by $\hat{\beta}_{en}$ is of utmost importance. Instead of measuring deviations from "across-the-curve" predictions, the median of the estimated regression coefficient $\hat{\beta}_{en}$ on μ and σ are presented in Tables A4 and A5. Coefficients estimated using 2SGAMLSS repeatedly match those of the benchmark model. In some circumstances the 2SGAMLSS fails to match the benchmark estimates, however increasing the size of the sample resolves this issue. This behaviour is observed on both the location and scale parameters. Estimation of treatment effects in the scale parameter only seemed challenging in scenario S7, although the precision in the location parameter remained superior compared to the other considered procedures. The estimates of x_{en}'s regression coefficient from the naive model either under- or overestimate the covariate's effect on both distributional parameters. The results from the binary endogeneity scenarios showcase the flexibility of the distributional regression framework. It allows us to accommodate different types of covariate effects without compromising the accuracy of the model coefficients.

The bias of the treatment effects on the mean or standard deviation estimated using 2SGAMLSS exhibit a lower relative bias than those obtained using different IV methods. These results indicate that our proposed estimation approach is able to recover treatment effects on different distributional quantities.

A.2.3 Bootstrap confidence intervals

Table A12 shows the coverage probabilities of the endogenous regressor's true effect obtained from the bootstrap confidence intervals for 2SGAMLSS outlined in Section 3.4 using $N_b = N_d = 100$. The coverage of the 2SGAMLSS confidence intervals vary depending on the considered scenario. As shown in the main paper in Table 2, the intervals produce satisfactory coverage probabilities, whereas for more complex effect forms of x_{en} (S1-S4) some under-coverage might be experienced. The coverage probabilities for ϑ_2 in scenario S2 are missing due to the considered response distribution being the Bernoulli distribution. We recommend to carefully select the bootstrap parameters N_d & N_b in order to obtain intervals that preserve their nominal coverage probabilities.

Table A2: Precision of (continuous) x_{en}'s estimated effect on the location parameter ϑ_1 across different sample sizes and scenarios.

	Mean bias		Median bias		IQR		RMSE	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
Scenario S1: Gaussian x_{en} ~ Gaussian y								
				$N = 500$				
naive	0.27	0.47	0.24	0.35	0.23	0.26	0.37	0.61
2SLS	0.98	2.02	0.81	1.64	0.99	2.09	1.04	2.27
2SRI	0.27	0.85	0.22	0.69	0.29	0.81	0.34	1.03
2SGAM	0.38	0.57	0.36	0.53	0.35	0.45	0.57	0.82
2SGAMLSS	0.41	0.50	0.23	0.37	0.26	0.31	0.59	0.70
benchmark	0.48	0.34	0.20	0.28	0.24	0.23	0.64	0.48
				$N = 2000$				
naive	0.18	0.29	0.18	0.29	0.11	0.14	0.23	0.37
2SLS	0.62	1.02	0.55	0.85	0.62	1.10	0.65	1.15
2SRI	0.13	0.43	0.11	0.34	0.13	0.38	0.16	0.53
2SGAM	0.18	0.31	0.17	0.28	0.17	0.20	0.27	0.44
2SGAMLSS	0.12	0.29	0.10	0.20	0.12	0.11	0.16	0.38
benchmark	0.11	0.19	0.09	0.17	0.10	0.07	0.15	0.27
				$N = 4000$				
naive	0.18	0.28	0.18	0.28	0.08	0.10	0.22	0.35
2SLS	0.56	0.77	0.51	0.64	0.50	0.75	0.58	0.86
2SRI	0.09	0.33	0.08	0.26	0.10	0.26	0.11	0.41
2SGAM	0.13	0.23	0.12	0.21	0.13	0.13	0.19	0.33
2SGAMLSS	0.08	0.18	0.07	0.17	0.08	0.07	0.11	0.25
benchmark	0.07	0.16	0.06	0.15	0.08	0.06	0.10	0.22
Scenario S2: Gaussian x_{en} ~ Binomial y								
				$N = 500$				
naive	0.24	0.24	0.18	0.23	0.08	0.07	0.31	0.30
2SLS	0.16	0.21	0.16	0.20	0.02	0.04	0.20	0.27
2SRI	0.13	0.35	0.11	0.29	0.13	0.28	0.15	0.44
2SGAM	0.17	0.25	0.16	0.22	0.15	0.13	0.24	0.35
2SGAMLSS	0.12	0.19	0.10	0.17	0.11	0.07	0.16	0.27
benchmark	0.16	0.18	0.10	0.17	0.09	0.05	0.21	0.25
				$N = 2000$				

Table A2: Precision of (continuous) x_{en}'s estimated effect on the location parameter ϑ_1 across different sample sizes and scenarios.

	Mean bias		Median bias		IQR		RMSE	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	0.17	0.20	0.17	0.20	0.04	0.04	0.21	0.25
2SLS	0.15	0.17	0.14	0.17	0.01	0.02	0.18	0.23
2SRI	0.06	0.23	0.06	0.19	0.07	0.11	0.08	0.28
2SGAM	0.08	0.13	0.07	0.12	0.07	0.07	0.10	0.18
2SGAMLSS	0.09	0.12	0.07	0.11	0.06	0.05	0.11	0.15
benchmark	0.07	0.12	0.07	0.11	0.05	0.04	0.09	0.15
				$N = 4000$				
naive	0.16	0.19	0.17	0.19	0.03	0.03	0.20	0.23
2SLS	0.14	0.17	0.14	0.17	0.01	0.01	0.17	0.22
2SRI	0.05	0.20	0.04	0.17	0.05	0.07	0.06	0.25
2SGAM	0.06	0.10	0.05	0.09	0.05	0.05	0.07	0.12
2SGAMLSS	0.06	0.09	0.06	0.09	0.05	0.04	0.08	0.12
benchmark	0.06	0.10	0.06	0.09	0.04	0.03	0.08	0.12
Scenario S3: Gaussian x_{en} ~ Negative Binomial y								
				$N = 500$				
naive	0.20	0.28	0.20	0.28	0.05	0.06	0.25	0.33
2SLS	2.38	1.75	2.31	1.45	1.51	1.79	2.51	2.02
2SRI	0.06	0.28	0.05	0.25	0.07	0.13	0.08	0.35
2SGAM	0.07	0.17	0.06	0.16	0.06	0.09	0.09	0.22
2SGAMLSS	0.07	0.13	0.07	0.12	0.06	0.07	0.09	0.17
benchmark	0.07	0.13	0.06	0.12	0.06	0.06	0.08	0.16
				$N = 2000$				
naive	0.20	0.26	0.20	0.26	0.03	0.04	0.24	0.30
2SLS	2.40	1.00	2.39	0.84	0.86	1.00	2.51	1.17
2SRI	0.03	0.25	0.03	0.23	0.04	0.09	0.04	0.30
2SGAM	0.04	0.08	0.03	0.07	0.03	0.04	0.04	0.11
2SGAMLSS	0.04	0.07	0.03	0.06	0.03	0.03	0.05	0.09
benchmark	0.03	0.07	0.03	0.06	0.03	0.03	0.04	0.09
				$N = 4000$				
naive	0.20	0.26	0.20	0.26	0.02	0.03	0.24	0.29
2SLS	2.38	0.79	2.35	0.69	0.57	0.76	2.49	0.93
2SRI	0.02	0.24	0.02	0.23	0.02	0.07	0.03	0.29
2SGAM	0.02	0.06	0.02	0.05	0.02	0.03	0.03	0.08
2SGAMLSS	0.03	0.05	0.02	0.05	0.02	0.02	0.03	0.07
benchmark	0.02	0.05	0.02	0.04	0.02	0.02	0.03	0.06
Scenario S4: Gaussian x_{en} ~ Gamma y								
				$N = 500$				
naive	3.19	3.96	0.20	0.26	0.07	0.09	3.92	5.03
2SLS	0.63	0.52	0.58	0.41	0.48	0.46	0.64	0.63
2SRI	0.07	0.27	0.06	0.23	0.08	0.13	0.09	0.34

Table A2: Precision of (continuous) x_{en}'s estimated effect on the location parameter ϑ_1 across different sample sizes and scenarios.

	Mean bias		Median bias		IQR		RMSE	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
2SGAM	0.08	0.17	0.07	0.17	0.08	0.09	0.10	0.22
2SGAMLSS	4.62	3.80	0.10	0.16	2.99	0.15	6.39	4.50
benchmark	1.65	2.29	0.06	0.13	0.06	0.08	2.12	2.75
				$N = 2000$				
naive	0.89	1.45	0.19	0.23	0.03	0.04	1.11	1.82
2SLS	0.58	0.31	0.57	0.24	0.27	0.17	0.57	0.38
2SRI	0.04	0.22	0.03	0.20	0.04	0.08	0.05	0.27
2SGAM	0.04	0.09	0.04	0.09	0.04	0.04	0.05	0.12
2SGAMLSS	2.17	1.13	0.04	0.08	0.04	0.04	2.56	1.11
benchmark	0.38	0.29	0.03	0.08	0.03	0.04	0.47	0.33
				$N = 4000$				
naive	0.77	1.17	0.19	0.23	0.02	0.03	0.93	1.49
2SLS	0.57	0.25	0.57	0.21	0.20	0.09	0.57	0.33
2SRI	0.03	0.21	0.02	0.20	0.03	0.06	0.03	0.26
2SGAM	0.03	0.08	0.03	0.07	0.03	0.03	0.04	0.10
2SGAMLSS	1.36	1.35	0.03	0.07	0.03	0.03	2.05	1.55
benchmark	0.15	0.20	0.02	0.06	0.02	0.03	0.21	0.28

Table A3: Precision of (continuous) x_{en}'s estimated effect on ϑ_2 across different sample sizes and scenarios.

	Scenario S1: Gaussian x_{en} ~ Gaussian y							
	Mean bias		Median bias		IQR		RMSE	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
				$N = 500$				
naive	1.53	0.57	0.18	0.14	0.04	0.26	2.24	0.76
2SLS	34.92	53.36	34.57	52.61	5.39	2.09	34.92	53.36
2SRI	34.75	51.79	34.36	51.48	5.27	0.45	34.76	51.79
2SGAM	34.15	50.72	33.86	50.37	5.36	0.81	34.15	50.72
2SGAMLSS	0.84	1.21	0.05	0.08	0.05	0.31	1.15	1.69
benchmark	1.95	0.33	0.04	0.06	0.04	0.23	3.00	0.45
				$N = 2000$				
naive	0.17	0.13	0.17	0.13	0.11	0.14	0.21	0.16
2SLS	34.10	52.01	33.98	51.90	0.62	1.10	34.10	52.01
2SRI	34.01	51.67	33.89	51.55	0.17	0.20	34.01	51.67
2SGAM	33.86	51.28	33.73	51.17	0.13	0.38	33.86	51.29
2SGAMLSS	0.04	0.07	0.02	0.04	0.12	0.11	0.05	0.10
benchmark	0.02	0.03	0.02	0.03	0.10	0.07	0.03	0.04
				$N = 4000$				
naive	0.17	0.13	0.17	0.13	0.08	0.10	0.20	0.15

2SLS	33.72	51.68	33.74	51.67	0.50	0.75	33.72	51.68
2SRI	33.64	51.48	33.64	51.43	0.13	0.13	33.64	51.48
2SGAM	33.57	51.20	33.57	51.18	0.10	0.26	33.57	51.20
2SGAMLSS	0.02	0.03	0.02	0.03	0.08	0.07	0.02	0.04
benchmark	0.01	0.03	0.01	0.02	0.08	0.06	0.02	0.04
Scenario S3: Gaussian x_{en} ∼ Negative Binomial y								
				$N = 500$				
naive	0.24	0.49	0.23	0.48	0.21	0.22	0.30	0.59
2SLS	5.47	6.26	5.45	6.18	0.60	0.75	5.48	6.26
2SRI	1.01	0.75	1.00	0.74	0.11	0.08	1.03	0.82
2SGAM	1.00	0.89	1.00	0.88	0.12	0.12	1.02	0.95
2SGAMLSS	0.24	0.23	0.20	0.18	0.24	0.20	0.29	0.28
benchmark	0.23	0.25	0.20	0.20	0.23	0.22	0.28	0.31
				$N = 2000$				
naive	0.24	0.52	0.24	0.51	0.11	0.13	0.29	0.63
2SLS	5.45	6.25	5.44	6.24	0.30	0.39	5.45	6.26
2SRI	1.00	0.74	1.00	0.74	0.06	0.04	1.02	0.81
2SGAM	1.00	0.88	1.00	0.88	0.05	0.06	1.02	0.94
2SGAMLSS	0.17	0.21	0.14	0.19	0.17	0.17	0.20	0.26
benchmark	0.17	0.21	0.15	0.19	0.17	0.17	0.20	0.26
				$N = 4000$				
naive	0.24	0.52	0.24	0.52	0.08	0.10	0.29	0.63
2SLS	5.46	6.25	5.45	6.24	0.22	0.29	5.47	6.26
2SRI	0.99	0.74	1.00	0.74	0.04	0.03	1.01	0.81
2SGAM	0.99	0.88	1.00	0.88	0.04	0.04	1.01	0.93
2SGAMLSS	0.16	0.22	0.15	0.22	0.14	0.14	0.20	0.28
benchmark	0.16	0.22	0.15	0.21	0.13	0.13	0.20	0.28
Scenario S4: Gaussian x_{en} ∼ Gamma y								
				$N = 500$				
naive	1.64	1.94	0.23	0.30	0.11	0.11	1.99	2.28
2SLS	1.51	1.70	1.45	1.65	0.35	0.31	1.52	1.73
2SRI	1.51	1.86	1.43	1.77	0.40	0.44	1.52	1.89
2SGAM	1.45	1.51	1.40	1.47	0.37	0.29	1.46	1.54
2SGAMLSS	3.11	2.23	0.26	0.15	1.27	0.21	3.81	2.62
benchmark	1.28	1.42	0.24	0.12	0.12	0.13	1.57	1.68
				$N = 2000$				
naive	0.47	0.77	0.23	0.29	0.04	0.04	0.56	0.90
2SLS	1.47	1.68	1.45	1.67	0.21	0.17	1.48	1.70
2SRI	1.49	1.94	1.47	1.90	0.25	0.28	1.50	1.96
2SGAM	1.47	1.61	1.45	1.59	0.24	0.20	1.48	1.64
2SGAMLSS	0.88	0.44	0.17	0.10	0.08	0.08	1.07	0.52
benchmark	0.40	0.15	0.17	0.09	0.05	0.06	0.49	0.19
				$N = 4000$				
naive	0.39	0.64	0.24	0.28	0.03	0.03	0.46	0.74
2SLS	1.46	1.66	1.44	1.65	0.14	0.14	1.46	1.69

2SRI	1.48	1.95	1.46	1.92	0.20	0.24	1.48	1.97
2SGAM	1.46	1.65	1.44	1.63	0.20	0.17	1.47	1.68
2SGAMLSS	0.73	0.59	0.15	0.09	0.06	0.06	0.89	0.70
benchmark	0.17	0.14	0.15	0.09	0.04	0.05	0.21	0.17

Table A4: Median $\hat{\beta}_{en}$ on the location parameter ϑ_1 across different sample sizes and scenarios using $C = 1000$ Monte Carlo replications.

	Scenario S5: Bernoulli $x_{en} \sim$ Gaussian y					
	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	1.276	0.657	1.268	0.632	1.274	0.645
2SLS	0.994	1.082	0.973	0.992	1.009	1.018
2SRI	0.973	1.062	0.972	0.968	1.010	1.021
2SGAM	0.979	1.038	0.966	1.002	1.018	1.052
2SGAMLSS	0.968	0.985	0.943	1.032	0.997	1.072
benchmark	0.973	0.984	1.001	0.992	1.013	0.977
	Scenario S6: Bernoulli $x_{en} \sim$ Bernoulli y					
naive	0.780	0.384	0.760	0.380	0.760	0.380
2SLS	0.169	0.149	0.155	0.151	0.154	0.155
2SRI	1.099	0.905	1.009	0.918	1.005	0.927
2SGAM	1.108	0.883	0.999	0.984	0.988	0.986
2SGAMLSS	0.633	0.565	0.568	0.627	0.569	0.636
benchmark	0.617	0.603	0.603	0.607	0.610	0.604
	Scenario S7: Bernoulli $x_{en} \sim$ Negative Binomial y					
naive	1.258	0.598	1.263	0.610	1.264	1.689
2SLS	5.014	4.501	4.972	4.705	4.969	4.747
2SRI	1.041	0.938	1.015	1.006	1.015	0.839
2SGAM	1.034	0.931	1.014	0.985	1.016	0.788
2SGAMLSS	0.972	0.959	0.959	1.022	0.947	1.437
benchmark	1.007	0.998	1.007	1.002	1.003	1.707
	Scenario S8: Bernoulli $x_{en} \sim$ Gamma y					
naive	0.514	0.428	0.490	0.424	0.485	0.424
2SLS	0.762	0.712	0.729	0.677	0.711	0.667
2SRI	0.733	0.693	0.696	0.664	0.686	0.671
2SGAM	0.709	0.657	0.698	0.668	0.687	0.667
2SGAMLSS	0.730	0.682	0.712	0.686	0.705	0.689
benchmark	0.745	0.689	0.706	0.671	0.698	0.664
	Scenario S9: Bernoulli $x_{en} \sim$ Logistic y					
naive	1.291	0.674	1.236	0.642	1.277	0.657
2SLS	1.079	0.895	0.923	1.098	1.009	1.068
2SRI	1.028	0.816	0.946	1.098	1.007	1.058
2SGAM	1.125	1.032	0.938	0.954	0.965	1.118
2SGAMLSS	1.188	1.044	0.997	0.994	0.963	1.040
benchmark	1.099	1.069	0.958	0.976	0.996	0.998

[a] The 2SLS procedure is fitted on a Gaussian distribution on all displayed scenarios (S5-S9), i.e. this estimator is mis-specified in scenarios S6-S9.
[b] 2SRI is fitted in scenario S7 using the `R` package `MASS` (Venables and Ripley, 2002).

Table A5: Median $\hat{\beta}_{en}$ on the scale parameter ϑ_2 across different sample sizes and scenarios using $C = 1000$ Monte Carlo replications.

	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
Scenario S5: Bernoulli $x_{en} \sim$ Gaussian y						
naive	1.269	1.215	1.261	1.210	1.260	1.204
2SGAMLSS	0.959	1.005	0.940	0.967	0.940	0.955
benchmark	1.010	1.009	1.000	1.005	1.000	0.999
Scenario S7: Bernoulli $x_{en} \sim$ Negative Binomial y						
naive	0.959	1.005	0.940	0.967	0.940	0.955
2SGAMLSS	1.010	1.009	1.000	1.005	1.000	0.999
benchmark	1.269	1.215	1.261	1.210	1.260	1.204
Scenario S8: Bernoulli $x_{en} \sim$ Gamma y						
naive	1.020	1.590	0.967	1.569	0.947	1.565
2SGAMLSS	1.463	1.268	1.405	1.230	1.383	1.224
benchmark	1.512	1.368	1.421	1.344	1.396	1.337
Scenario S9: Bernoulli $x_{en} \sim$ Logistic y						
naive	1.277	1.232	1.270	1.221	1.273	1.219
2SGAMLSS	0.984	1.009	0.946	0.969	0.944	0.967
benchmark	1.007	1.014	0.999	1.003	1.002	1.001

Table A6: Median relative bias of the estimated marginal effect at means (MEM) on the mean (continuous treatment)

	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
Scenario S1: Gaussian x_{en} ~ Gaussian y						
naive	-1.067	1.050	-1.101	0.743	-1.163	0.679
2SLS	-0.053	0.956	0.000	0.973	-0.014	0.965
2SRI	-0.055	0.951	-0.006	0.977	-0.016	0.965
2SGAM	-0.113	0.615	0.002	0.211	-0.025	0.158
2SGAMLSS	-0.131	0.596	-0.012	0.228	-0.041	0.162
Benchmark	-0.053	0.616	0.018	0.263	-0.030	0.219
Scenario S2: Gaussian x_{en} ~ Bernoulli y						
naive	1.085	1.113	1.203	0.719	1.224	0.644
2SLS	-0.042	1.001	-0.091	0.996	-0.050	0.993
2SRI	0.023	1.006	-0.016	0.994	0.010	0.986
2SGAM	0.056	0.690	-0.005	0.235	0.012	0.141
2SGAMLSS	0.007	0.689	-0.020	0.262	-0.005	0.176
Benchmark	0.034	0.744	-0.024	0.329	0.001	0.240
Scenario S3: Gaussian x_{en} ~ Negative Binomial y						
naive	1.166	0.933	1.235	0.834	1.292	0.814
2SLS	0.071	1.067	0.032	1.126	0.049	1.135
2SRI	0.032	1.142	-0.012	1.186	0.000	1.190
2SGAM	0.098	0.512	0.005	0.205	0.014	0.153
2SGAMLSS	0.083	0.307	0.007	0.154	0.012	0.144
Benchmark	0.007	0.308	-0.002	0.171	0.005	0.116
Scenario S4: Gaussian x_{en} ~ Gamma y						
naive	1.173	0.922	1.331	0.778	1.388	0.745
2SLS	0.109	1.082	0.043	1.051	0.023	1.062
2SRI	0.069	1.131	0.022	1.082	-0.003	1.083
2SGAM	0.145	0.624	0.041	0.176	0.019	0.168
2SGAMLSS	0.070	0.393	0.027	0.139	0.010	0.158
Benchmark	0.032	0.349	-0.004	0.191	-0.002	0.157

Table A7: Median relative bias of the estimated marginal effect at means (MEM) on the standard deviation (continuous treatment)

	Scenario S1: Gaussian x_{en} ~ Gaussian y					
	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	-1.031	-0.622	-1.135	-0.832	-1.167	-0.953
2SGAMLSS	0.017	0.080	-0.015	0.057	-0.005	0.071
Benchmark	0.024	0.057	-0.007	0.061	0.005	0.060
Scenario S2: Gaussian x_{en} ~ Bernoulli y						
naive	0.993	1.107	0.983	0.746	0.983	0.682
2SGAMLSS	1.097	0.673	1.090	0.256	1.084	0.171
Benchmark	1.092	0.710	1.090	0.333	1.085	0.246
Scenario S3: Gaussian x_{en} ~ Negative Binomial y						
naive	1.182	0.463	1.253	0.364	1.292	0.333
2SGAMLSS	0.097	0.274	0.006	0.153	0.009	0.146
Benchmark	0.028	0.259	-0.016	0.172	0.003	0.125
Scenario S4: Gaussian x_{en} ~ Gamma y						
naive	1.213	0.393	1.369	0.225	1.422	0.171
2SGAMLSS	0.035	0.332	-0.015	0.186	-0.020	0.196
Benchmark	0.027	0.332	0.001	0.201	-0.006	0.158

Table A8: Median relative bias of the estimated average marginal effect (AME) on the mean (binary treatment)

	Scenario S5: Bernoulli x_{en} ~ Gaussian y					
	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	-0.276	0.359	-0.272	0.362	-0.276	0.362
2SLS	0.016	-0.005	0.007	-0.028	-0.007	-0.003
2SRI	0.007	0.005	0.004	-0.021	-0.006	0.002
2SGAM	-0.006	0.005	0.000	-0.007	-0.010	-0.016
2SGAMLSS	0.025	-0.009	0.050	-0.037	0.021	-0.060
benchmark	0.012	0.017	0.004	0.015	-0.006	0.016
	Scenario S6: Bernoulli x_{en} ~ Bernoulli y					
naive	-0.291	0.356	-0.272	0.360	-0.273	0.359
2SLS	-0.118	0.025	-0.031	0.018	-0.025	-0.011
2SRI	-0.100	0.069	-0.012	0.056	-0.012	0.032
2SGAM	-0.094	0.110	-0.005	0.011	0.002	0.004
2SGAMLSS	-0.023	0.085	0.053	-0.032	0.053	-0.043
benchmark	-0.006	0.028	0.009	0.016	-0.006	0.014
	Scenario S7: Bernoulli x_{en} ~ Negative Binomial y					
naive	-0.270	0.452	-0.272	0.448	-0.272	0.451
2SLS	0.007	0.108	0.016	0.085	0.023	0.064
2SRI	-0.026	0.067	-0.016	0.022	-0.012	-0.005
2SGAM	-0.018	0.089	-0.015	0.029	-0.011	0.017
2SGAMLSS	0.033	0.030	0.037	-0.043	0.053	-0.065
benchmark	0.000	0.034	-0.000	0.016	0.001	0.016
	Scenario S8: Bernoulli x_{en} ~ Gamma y					
naive	0.263	0.387	0.289	0.402	0.295	0.401
2SLS	-0.033	0.006	-0.013	0.036	0.005	0.050
2SRI	-0.064	-0.073	-0.043	-0.001	-0.025	-0.014
2SGAM	-0.029	0.004	-0.045	-0.005	-0.026	-0.002
2SGAMLSS	-0.070	-0.046	-0.064	-0.029	-0.058	-0.044
benchmark	-0.113	-0.047	-0.052	-0.005	-0.046	0.003
	Scenario S9: Bernoulli x_{en} ~ Logistic y					
naive	-0.291	0.326	-0.236	0.358	-0.277	0.343
2SLS	-0.079	0.105	0.077	-0.098	-0.009	-0.068
2SRI	-0.028	0.184	0.054	-0.098	-0.007	-0.058
2SGAM	-0.125	-0.032	0.062	0.046	0.035	-0.118
2SGAMLSS	-0.180	-0.035	0.003	0.006	0.038	-0.040
benchmark	-0.099	-0.069	0.042	0.024	0.004	0.002

Table A9: Median relative bias of the estimated marginal effect at means (MEM) on the mean (binary treatment)

	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
Scenario S5: Bernoulli x_{en} ~ Gaussian y						
naive	-0.276	0.359	-0.272	0.362	-0.276	0.362
2SLS	0.016	-0.005	0.007	-0.028	-0.007	-0.003
2SRI	0.007	0.005	0.004	-0.021	-0.006	0.002
2SGAM	-0.006	0.005	0.000	-0.007	-0.010	-0.016
2SGAMLSS	0.025	-0.009	0.050	-0.037	0.021	-0.060
benchmark	0.012	0.017	0.004	0.015	-0.006	0.016
Scenario S6: Bernoulli x_{en} ~ Bernoulli y						
naive	-0.303	0.426	-0.277	0.418	-0.281	0.409
2SLS	-0.123	0.258	-0.035	0.253	-0.030	0.230
2SRI	-0.098	0.295	-0.013	0.281	-0.014	0.265
2SGAM	-0.093	0.172	-0.010	0.057	0.002	0.039
2SGAMLSS	-0.020	0.168	0.045	0.052	0.050	0.032
benchmark	-0.007	0.103	0.006	0.079	-0.010	0.063
Scenario S7: Bernoulli x_{en} ~ Negative Binomial y						
naive	-0.317	0.351	-0.329	0.366	-0.326	0.373
2SLS	-0.075	-0.734	-0.068	-0.775	-0.061	-0.819
2SRI	-0.054	-0.793	-0.049	-0.883	-0.044	-0.919
2SGAM	-0.053	-0.034	-0.050	-0.078	-0.047	-0.080
2SGAMLSS	0.019	-0.126	0.010	-0.183	0.031	-0.196
benchmark	0.006	-0.006	-0.001	0.006	0.000	0.014
Scenario S8: Bernoulli x_{en} ~ Gamma y						
naive	0.246	0.337	0.276	0.356	0.283	0.360
2SLS	-0.075	-0.517	-0.053	-0.465	-0.034	-0.445
2SRI	-0.069	-0.628	-0.054	-0.512	-0.036	-0.526
2SGAM	-0.027	-0.107	-0.047	-0.104	-0.039	-0.098
2SGAMLSS	-0.079	-0.126	-0.079	-0.103	-0.074	-0.106
benchmark	-0.090	-0.061	-0.051	-0.009	-0.038	0.003
Scenario S9: Bernoulli x_{en} ~ Logistic y						
naive	-0.291	0.326	-0.236	0.358	-0.277	0.343
2SLS	-0.079	0.105	0.077	-0.098	-0.009	-0.068
2SRI	-0.028	0.184	0.054	-0.098	-0.007	-0.058
2SGAM	-0.125	-0.032	0.062	0.046	0.035	-0.118
2SGAMLSS	-0.180	-0.035	0.003	0.006	0.038	-0.040
benchmark	-0.099	-0.069	0.042	0.024	0.004	0.002

Table A10: Median relative bias of the estimated average marginal effect (AME) on the standard deviation (binary treatment)

	Scenario S5: Bernoulli x_{en} ~ Gaussian y					
	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	-0.272	-0.225	-0.270	-0.231	-0.270	-0.219
2SGAMLSS	0.045	0.007	0.061	0.059	0.061	0.059
benchmark	-0.004	-0.008	0.001	-0.005	0.001	0.004
	Scenario S6: Bernoulli x_{en} ~ Bernoulli y					
naive	-0.282	0.341	-0.267	0.348	-0.270	0.348
2SGAMLSS	-0.030	0.061	0.043	-0.064	0.040	-0.079
benchmark	-0.011	0.003	-0.005	-0.017	-0.017	-0.017
	Scenario S7: Bernoulli x_{en} ~ Negative Binomial y					
naive	-0.301	0.084	-0.307	0.073	-0.309	0.082
2SGAMLSS	-0.042	-0.043	-0.009	-0.088	0.013	-0.091
benchmark	0.005	0.043	0.002	0.015	0.001	0.019
	Scenario S8: Bernoulli x_{en} ~ Gamma y					
naive	0.293	0.056	0.321	0.069	0.332	0.070
2SGAMLSS	-0.127	-0.065	-0.081	-0.022	-0.079	-0.026
benchmark	-0.145	-0.020	-0.081	-0.005	-0.058	-0.003
	Scenario S9: Bernoulli x_{en} ~ Logistic y					
naive	-0.305	-0.302	-0.301	-0.291	-0.303	-0.293
2SGAMLSS	0.029	-0.012	0.055	0.040	0.053	0.037
benchmark	-0.000	0.006	0.003	0.004	-0.001	0.003

Table A11: Median relative bias of the estimated marginal effect at means (MEM) on the standard deviation (binary treatment)

	Scenario S5: Bernoulli x_{en} ~ Gaussian y					
	$N = 500$		$N = 2000$		$N = 4000$	
	Linear	Non-lin.	Linear	Non-lin.	Linear	Non-lin.
naive	-0.275	-0.243	-0.273	-0.243	-0.274	-0.232
2SGAMLSS	0.039	-0.010	0.057	0.051	0.060	0.052
benchmark	-0.008	-0.003	-0.000	0.005	0.000	0.013
	Scenario S6: Bernoulli x_{en} ~ Bernoulli y					
naive	-0.247	0.370	-0.225	0.387	-0.229	0.391
2SGAMLSS	0.021	0.064	0.068	-0.047	0.066	-0.028
benchmark	-0.006	0.012	0.003	0.026	-0.004	0.049
	Scenario S7: Bernoulli x_{en} ~ Negative Binomial y					
naive	-0.358	-0.005	-0.360	0.004	-0.364	0.017
2SGAMLSS	-0.053	-0.131	-0.029	-0.183	-0.023	-0.185
benchmark	0.013	0.054	0.006	0.037	0.000	0.039
	Scenario S8: Bernoulli x_{en} ~ Gamma y					
naive	0.247	-0.049	0.270	-0.025	0.284	-0.012
2SGAMLSS	-0.201	-0.161	-0.157	-0.124	-0.143	-0.114
benchmark	-0.111	-0.025	-0.070	0.022	-0.046	0.026
	Scenario S9: Bernoulli x_{en} ~ Logistic y					
naive	-0.356	-0.414	-0.356	-0.403	-0.358	-0.404
2SGAMLSS	0.013	-0.070	0.033	-0.036	0.029	-0.022
benchmark	0.004	0.035	0.002	0.041	-0.003	0.043

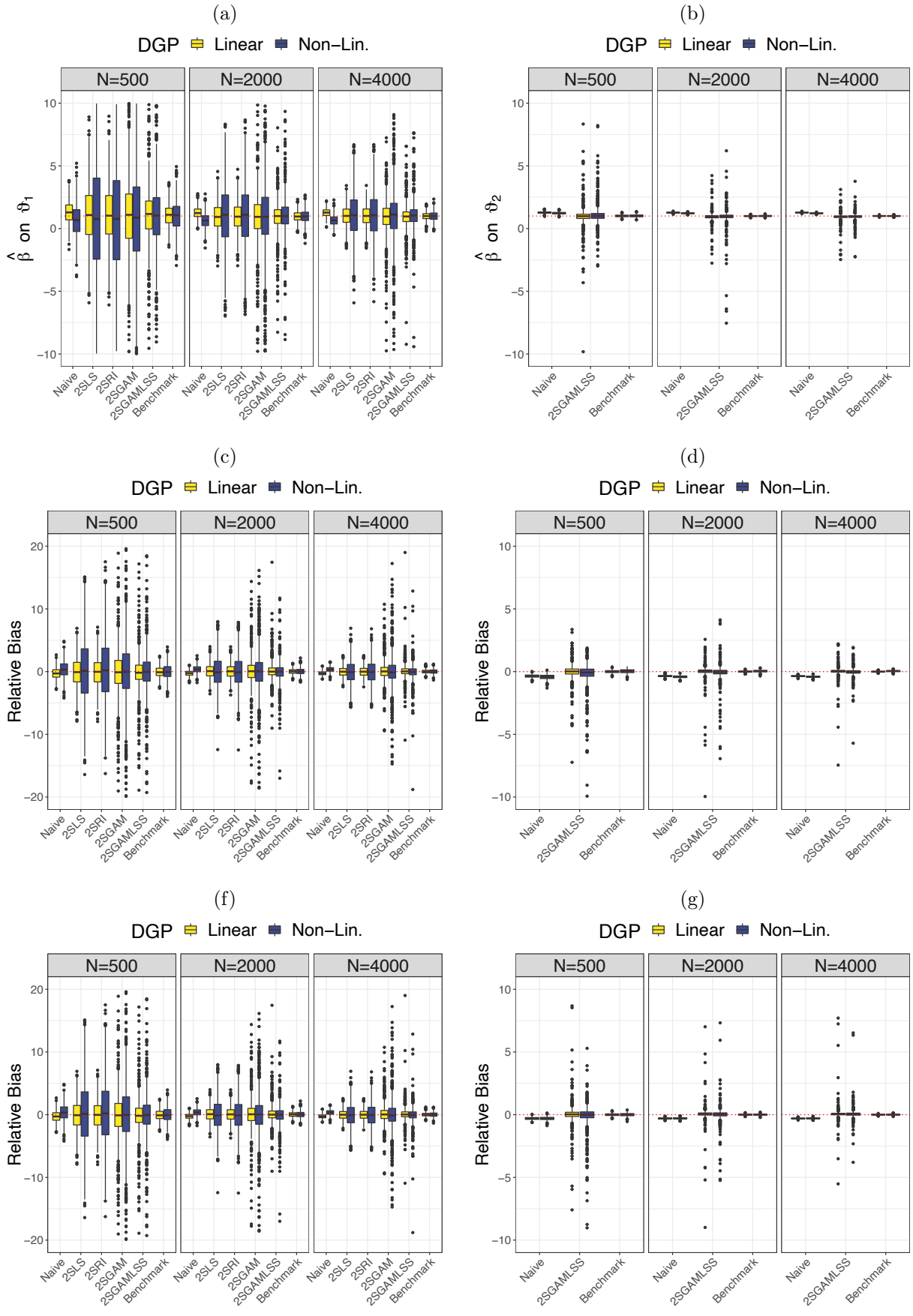

Figure A2: Boxplots of estimated coefficients on ϑ_1 (a) and ϑ_2 (b), as well as relative bias of MEM/AME on the mean (c)/(e), and standard deviation (d)/(f).

Table A12: Coverage probabilities of x_{en}'s true effect using 2SGAMLSS bootstrap confidence intervals ($N_b = N_d = 100$) across distributional parameters, scenarios, and sample sizes. $C = 200$ Monte Carlo replications.

	ϑ_1			ϑ_2		
	90%	95%	99%	90%	95%	99%
Scenario S1: Gaussian x_{en} ~ Gaussian y						
$N = 500$	0.91	0.96	0.99	0.51	0.60	0.74
$N = 2000$	0.77	0.85	0.94	0.51	0.60	0.74
$N = 4000$	1.00	1.00	1.00	0.51	0.56	0.65
Scenario S2: Gaussian x_{en} ~ Bernoulli y						
$N = 500$	0.72	0.80	0.89	-	-	-
$N = 2000$	0.63	0.71	0.82	-	-	-
$N = 4000$	0.57	0.64	0.76	-	-	-
Scenario S3: Gaussian x_{en} ~ Negative Binomial y						
$N = 500$	0.76	0.84	0.94	0.76	0.84	0.94
$N = 2000$	0.56	0.65	0.81	0.41	0.51	0.80
$N = 4000$	-	-	-	-	-	-
Scenario S4: Gaussian x_{en} ~ Gamma y						
$N = 500$	0.90	0.95	0.99	0.58	0.70	0.86
$N = 2000$	0.66	0.76	0.88	0.58	0.70	0.86
$N = 4000$	0.54	0.63	0.77	0.46	0.57	0.73
Scenario S5: Bernoulli x_{en} ~ Gaussian y						
$N = 500$	0.94	0.99	1.00	0.96	0.98	1.00
$N = 2000$	0.91	0.97	0.99	0.86	0.95	0.98
$N = 4000$	0.95	0.98	0.99	0.89	0.94	0.99
Scenario S6: Bernoulli x_{en} ~ Bernoulli y						
$N = 500$	0.96	0.99	1.00	-	-	-
$N = 2000$	0.96	0.98	1.00	-	-	-
$N = 4000$	0.94	0.97	1.00	-	-	-
Scenario S7: Bernoulli x_{en} ~ Negative Binomial y						
$N = 500$	0.98	1.00	1.00	0.51	0.67	0.95
$N = 2000$	0.94	0.99	1.00	0.28	0.46	0.81
$N = 4000$	-	-	-	-	-	-
Scenario S8: Bernoulli x_{en} ~ Gamma y						
$N = 500$	0.90	0.94	1.00	0.97	1.00	1.00
$N = 2000$	0.89	0.97	1.00	0.87	0.96	0.98
$N = 4000$	0.92	0.96	0.99	0.81	0.85	0.95
Scenario S9: Bernoulli x_{en} ~ Logistic y						
$N = 500$	0.93	0.97	1.00	0.99	0.99	1.00
$N = 2000$	0.94	0.97	1.00	0.91	0.96	0.99
$N = 4000$	0.89	0.96	1.00	0.97	1.00	1.00

[a] Scenarios S2 and S6 feature a response distribution with only one parameter (ϑ_1).

[b] Values displayed for S3 were obtained using $C = 100$ Monte Carlo replications.

[c] Values for $N = 4000$ are missing in scenarios S3 & S7 due to computation-time related constraints.

[d] Coverage probabilities for scenarios S1-S4 represent simultaneous "across-the-curve" coverage.

Table A13: Employed non-linear functions. All rescaled to lie in the domain [0,1].

Function	Definition
$f_1(x)$	$\cos(2\pi x)$
$f_2(x)$	$0.5\{x^3 + \sin(\pi x^3)\}$
$f_3(x)$	$\text{-}0.5\{x + \sin(\pi x^{2.5})\}$
$f_4(x)$	$\beta_{20,22}(x) - \beta_{4,10}(x) - \beta_{6,8}(x)$
$f_5(x)$	$x^{11}\{10(1-x)\}^6 + 10(10x)^3(1-x)^{10}$
$f_6(x)$	$\sin(x)^5 + (x^2-2)^9 + 10$
$f_7(x)$	$-\{6\beta_{18,12}(x) + 3\beta_{3,12}(x) + 5\beta_{3,12}(x)\} - 1$
$f_{11}(x)$	$\sin(2\pi x)$
$\beta_{l,m}(x)$	$\{\Gamma(l+m)/\Gamma(l)\Gamma(m)\}x^{l-1}(1-x)^{m-1}$

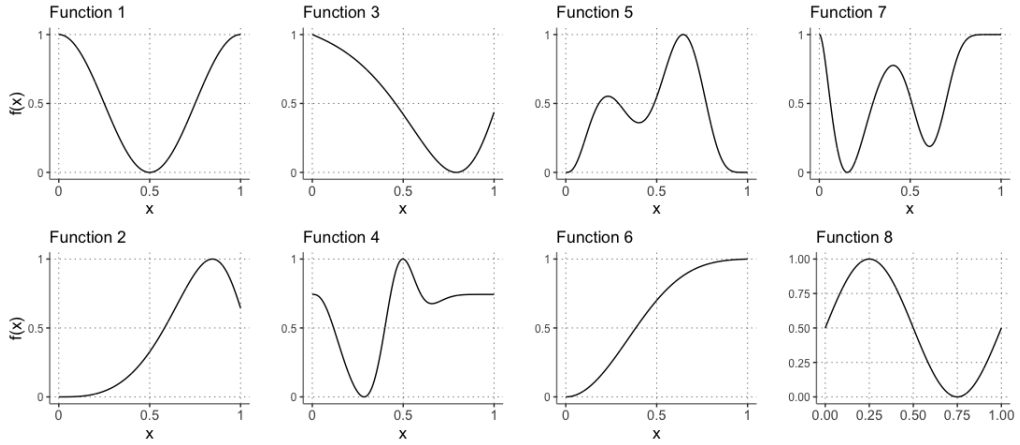

Figure A3: Employed non-linear functions.

B Additional figures of the application study

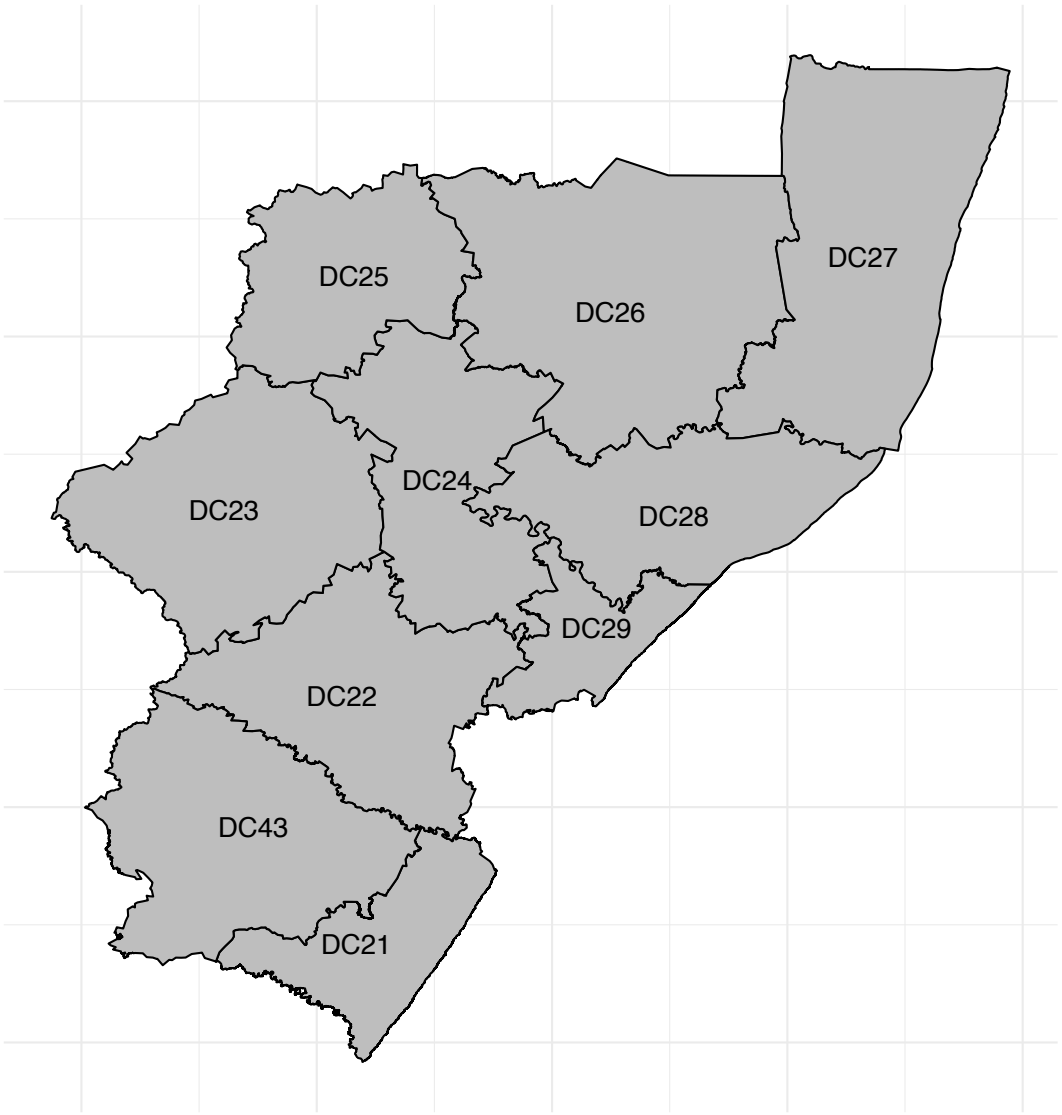

Figure B1: Map of the KwaZulu-Natal province with its 10 districts.

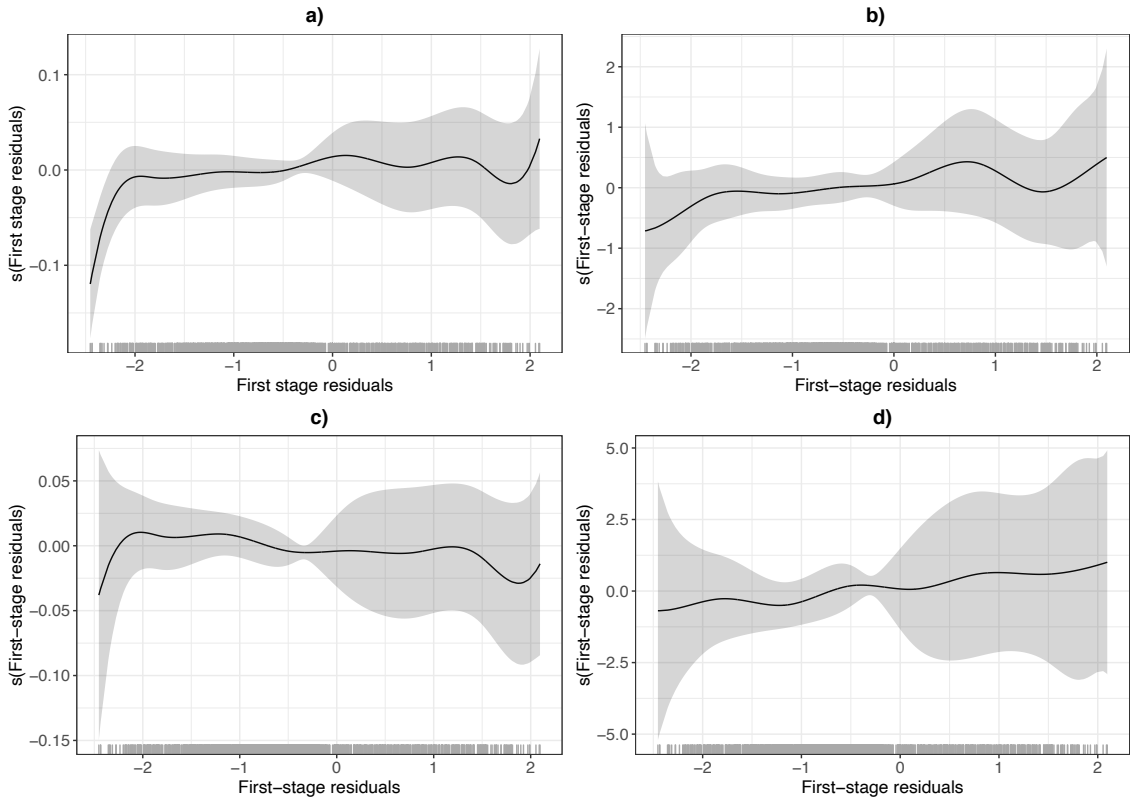

Figure B2: Estimated smooth effect of the first-stage residuals $\hat{\boldsymbol{\xi}}$ with 95% confidence bands on the predictors of ϑ_1 (a), and ϑ_2 (b) of Δ_t`prop_male_emp`. Estimated smooth effect of the first-stage residuals $\hat{\boldsymbol{\xi}}$ with 95% confidence bands on the predictors of ϑ_1 (c), and ϑ_2 (d) of Δ_t`prop_female_emp`.

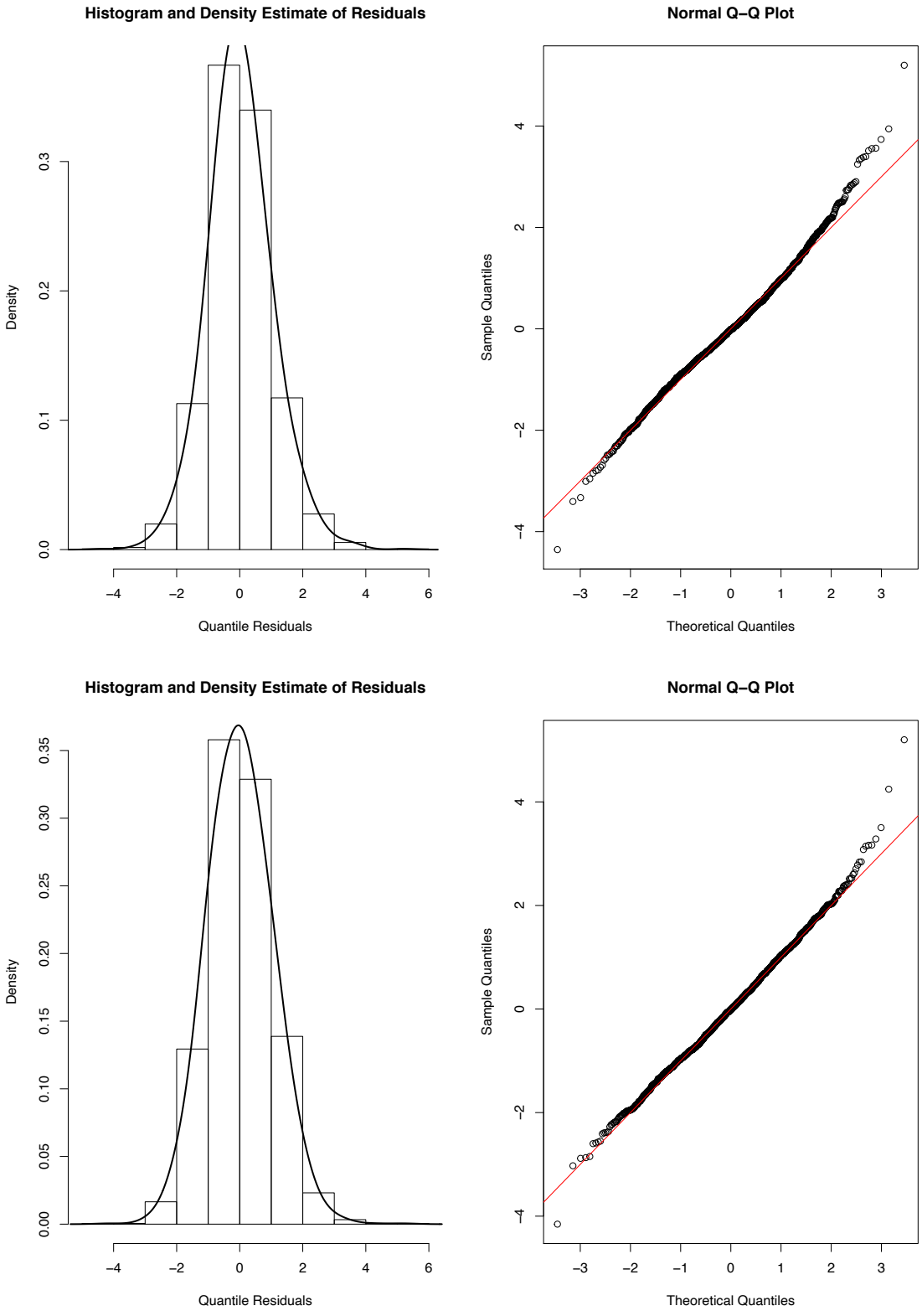

Figure B3: Quantile-Residual plots for a 2SGAMLSS of Δ_t`prop_female_emp` (top) and Δ_t`prop_male_emp` (bottom) assuming a logistic distribution.

Appendix B: Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses (with Supplement)

Joint work with Andreas Groll.

Briseño Sanchez, G. & Groll, A. (2024) Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses. [Manuscript submitted for publication.]

Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses

Guillermo Briseño Sanchez[1] and Andreas Groll[2]

[1]Methods for Big Data, Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany,

[2]Statistical Methods for Big Data, Department of Statistics, TU Dortmund University, Dortmund, Germany.

Abstract

We propose a distributional copula regression modelling approach for bivariate responses comprised of non-commensurate (i.e. mixed) variables. In our case, the margins are a right-censored time-to-event outcome and a non-time-to-event variable. The underlying hazard rate of the time-to-event margin is modelled using discrete-time-to-event or piecewise-exponential methods. A flexible statistical model is achieved by relying on the correspondence of the likelihood of the aforementioned time-to-event approaches with well-known univariate distributions. We construct joint bivariate distributions for these mixed responses by means of parametric bivariate copulas. This allows for separate specification of the dependence structure between the margins and their individual distribution functions. All coefficients of the distributional copula regression models considered here are estimated simultaneously via penalised maximum likelihood. We showcase the versatility of our proposed approach in an analysis of red-light running behaviour of E-cyclists by modelling the joint distribution of a mixed response comprised of a binary response and a time-to-event outcome that indicates the time of red traffic light running.

Keywords: Copula regression; Dependence modelling; GAMLSS; Semi-parametric regression; Time-to-event analysis.

1 Introduction

In many applications one may be interested in modelling the joint statistical behaviour of two random variables, say Y and T, instead of just one of them. Separate analyses of the individual outcomes are prone to miss important characteristics of their interrelationship. Parametric bivariate distributions are attractive candidates since they provide an interpretable closed form of the joint behaviour of interest. For example, consider continuous variables Y and T quantifying children's malnutrition via stunting and wasting scores. In this case, the bivariate Gaussian distribution with correlation parameter becomes a natural choice, see Klein et al. (2014). For discrete Y and T, as in the number of goals in football matches, the bivariate Poisson distribution with strictly positive covariance parameter could be appropriate, see Karlis and Ntzoufras (2003) as well as Groll et al. (2018). While there exists a vast literature on bivariate distributions and their applications, see e.g. Lai and Balakrishnan (2009) and Kocherlakota and Kocherlakota (2017), cases where they lose their appeal occur regularly. If the margins follow different families (e.g. Y Gaussian and T Gamma) and there is also interest in more complex dependence structures, finding a suitable bivariate distribution can become a challenging task. This problem is exacerbated when the marginal variables have different supports, e.g. Y continuous and T binary or discrete. Such non-commensurate or mixed outcomes are ubiquitous in health and medicine related applications (de Leon and Wu, 2011), but they are likely to be prevalent in other areas of research too. For example in transportation (Spissu et al., 2009), social sciences (Wagner and Tüchler, 2013), economics (Hohberg et al., 2021) and environmental sciences (Najib et al., 2022). Lastly, if one of the margins is subject to censoring (i.e. the actual response realisation is not observed), then the aforementioned bivariate distributions cease to be suitable tools for constructing a statistical model.

To sidestep these limitations, we resort to parametric bivariate copulas in order to construct joint bivariate distributions with arbitrary (albeit suitable) margins and a separate specification of their dependence structure. Therefore, we aim to provide a tool to analyse complex phenomena via a single statistical model of the multiple dimensions that constitute it. Our main interest is to model jointly bivariate outcomes that are non-commensurable or mixed, that is Y and T which have different supports and in addition, the latter is subject to censoring. Without loss of generality, for the remainder of this manuscript we denote the first margin Y as *non-time-to-event* and the second margin T as *time-to-event* variables, respectively. We propose to model the underlying hazard rate of the time-to-event margin T using discrete time (see Tutz and Schmid, 2016, DT) and piecewise-exponential (Bender et al., 2018, PW) methods within a distributional copula regression framework, such that a highly flexible statistical model can be fitted to the response vector. This is achieved by exploiting the likelihood of the aforementioned time-to-event approaches, which correspond to well-known likelihood functions of univariate Bernoulli (DT) and Poisson (PW) random variables, respectively.

Statistical models that are most closely related to this approach are *joint models* for longitudinal and time-to-event outcomes (Rizopoulos, 2012) and *landmarking* models (Suresh et al., 2019). On one hand, joint models aim to study the distribution of a longitudinal marker and a (potentially censored) time-to-event variable. On the other hand, landmarking does not model the entire (longitudinal) history of the non-time-to-event margin and instead conditions the joint distribution on subjects still not having experienced the event in a current observation time period. The two aforementioned methods differ from our contribution in some aspects. One concerns the non-time-to-event response in the sense that we do not work with longitudinal data. Another aspect is that for joint models the longitudinal margin, broadly speaking, receives most of the attention, whereas our focus is equally on both time-to-event and non-time-to-event margins. Nevertheless, our contribution could be employed for example in landmarking in the presence of markers that may be continuous, binary or discrete, see Suresh et al. (2019) and Suresh et al. (2021) for continuous and binary markers, respectively. Most recently, the likelihood of *DT* models has been used in joint models with a discrete-time component (Medina-Olivares et al., 2023), whereas Rappl et al. (2023) used the likelihood of a *PW* approach for joint models as well. Related work in the realm of bivariate distributional copula regression models is Marra and Radice (2017, bivariate continuous margins). Moreover, for mixed outcomes see Klein et al. (2018, binary and continuous margins) and Marra et al. (2020, binary and discrete margins), respectively.

The remainder of this manuscript is structured as follows: Section 2 reviews bivariate distributional copula regression and the considered time-to-event analysis methods. Our contribution is presented in Section 3. In Section 4, we apply our proposed method to analyse red-light running behaviour of E-cyclists using data from China (Gao et al., 2020). This is accomplished by modelling the joint distribution of a bivariate mixed outcome that consists of a binary non-time-to-event and a time-to-event margin, as well as their dependence. Lastly, a discussion is given in Section 5.

2 Methodology

In this section we introduce distributional regression for cases in which the joint distribution of a bivariate response is constructed by means of parametric copulas. Afterwards, we present discrete time and piecewise-exponential time-to-event models from a distributional regression perspective. The methods shown here apply for the i-th observation in a sample of size n, i.e. $i = 1, \ldots, n$. We omit this index in order to avoid clutter in the notation.

2.1 Bivariate distributional copula regression

The response is assumed to be an independent random draw from a parametric distribution. Our goal is to construct a model for the joint probability of the non-time-to-event variable and the time-to-event margin (occurring after t). In the context of bivariate outcomes with distributions constructed by means of copulas, this may be written as

$$P(Y \leq y,\ T > t;\ \boldsymbol{\vartheta}) = C[F_1(y;\ \boldsymbol{\vartheta}^{(1)}),\ S_2(t;\ \boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}], \tag{1}$$

where $C : (0,1)^2 \to (0,1)$ is a parametric copula function that connects or binds the margins, but does not depend on them (Nelsen, 2006). We do not assume that the true copula function is necessarily known in advance. The scalar quantity $\vartheta^{(c)}$, commonly known as "dependence parameter", determines the dependence structure between the margins. The range of the dependence parameter typically depends on the assumed or chosen copula function. However, it is possible to express the dependence between the margins in terms of Kendall's τ based on values of $\vartheta^{(c)}$, see Table D2 in Supplement D for the corresponding transformations. The margin-specific parameter vectors $\boldsymbol{\vartheta}^{(\bullet)} = (\vartheta_1^{(\bullet)}, \ldots, \vartheta_{K_{(\bullet)}}^{(\bullet)})^\top$ contain the $k = 1, \ldots, K_{(\bullet)}$, with $\bullet \in \{1, 2\}$, parameters that completely specify the distribution of the margins. In turn, the bivariate distribution depends on the vector $\boldsymbol{\vartheta} = ((\boldsymbol{\vartheta}^{(1)})^\top, (\boldsymbol{\vartheta}^{(2)})^\top, \vartheta^{(c)})^\top$, which contains all parameters of the margins and the dependence parameter. Following the Generalised Additive Models for Location, Scale and Shape (Stasinopoulos et al., 2017, GAMLSS) approach, potentially all parameters of the joint distribution are modelled as functions of the covariates at hand $\mathbf{x} = (x_1, x_2, \ldots, x_P)^\top \in \mathbb{R}^P$ by means of a structured additive predictor $\eta_k^{(\bullet)}$ in the following fashion:

$$\vartheta_k^{(\bullet)} = g_{(\bullet)k}^{-1}(\eta_k^{(\bullet)}) \ \Leftrightarrow\ g_{(\bullet)k}(\vartheta_k^{(\bullet)}) = \eta_k^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} s_r^{(\bullet)}(x_r),\ \bullet \in \{1, 2, c\}, \tag{2}$$

where $g^{-1}(\cdot)$ is a response function with corresponding inverse or "link" $g(\cdot)$, guaranteeing that the distribution parameters lie in their respective range. The range of the dependence parameter $\vartheta^{(c)}$ depends on the copula. The summation limit $P_k^{(\bullet)}$ emphasizes that the individual parameters must not be necessarily modelled using the same subset of covariates. The coefficient $\beta_{0k}^{(\bullet)}$ denotes a parameter-specific intercept and $s_r^{(\bullet)}(\cdot)$ are smooth functions that can accommodate a wide range of functional forms of the covariates such as linear, non-linear (e.g. P-Splines, Eilers and Marx, 1996), random effects or spatial. The smooth functions can be represented as linear combinations of basis functions and unknown regression coefficients to be estimated: $\boldsymbol{\eta}_k^{(\bullet)} = \mathbf{Z}_k^{(\bullet)} \boldsymbol{\beta}_k^{(\bullet)}$, where $\mathbf{Z}_k^{(\bullet)}$ is a design matrix of suitable dimensions, $\boldsymbol{\beta}_k^{(\bullet)}$ is a vector of coefficients and $\bullet \in \{1, 2, c\}$. Hence, the model ultimately depends on the vector $\boldsymbol{\beta} = ((\boldsymbol{\beta}^{(1)})^\top, (\boldsymbol{\beta}^{(2)})^\top, (\boldsymbol{\beta}^{(c)})^\top)^\top$, which contains the coefficients corresponding to the models of the margins' parameters and the copula dependence parameter, respectively. Note that for some covariate effects the coefficients may be associated with a quadratic penalty and a

smoothing parameter in order to enforce smoothness of the estimated functions, see Wood (2017) for more details. Distributional copula regression models as in Equations (1)-(2) can be fitted using either penalised maximum likelihood (Marra and Radice, 2017) (as in our case), fully Bayesian inference (Umlauf and Kneib, 2018) or boosting (Hans et al., 2022). Here we estimate all unknown coefficients in $\boldsymbol{\beta}$ simultaneously using penalised maximum likelihood, and base our software implementation **entirely** on the `R` package `GJRM` (Marra and Radice, 2023).

2.2 Considered time-to-event analysis approaches

Recall that T denotes the event time and let T^{cens} denote a random censoring time. In the case of right-censored data one observes $\tilde{T} = \min\{T, T^{cens}\}$ and its censoring indicator $\delta = \mathbb{1}\{T \leq T^{cens}\}$. We assume that the censoring time T^{cens} is non-informative and independent of the event time T. Note that it is possible to relax the independent censoring assumption and allow for dependence between T and T^{cens}, although this will result in a more complex model than the one we present here, see Czado and Van Keilegom (2022) for more details.

In this manuscript, we consider discrete time-to-event (DT) and piecewise-exponential (PW) approaches to model the underlying hazard function of the right-censored time-to-event variable. Let $\mathcal{I}_j$ with $j = 1, \ldots, J$ denote an interval of finite length with respective left and right bounds κ_{j-1}, κ_j, and $\kappa_0 = 0 < \kappa_1 < \cdots < \kappa_J < \infty$. In the DT framework, T takes values in the set $\{1, 2, \ldots, J\}$. One may assume that the process generating the observed times is truly discrete or that there are J intervals $\mathcal{I}_j$ that "cut" or discretize the underlying continuous follow-up. The conditional discrete hazard is defined as the following conditional probability (Tutz and Schmid, 2016): $\lambda(t \mid \mathbf{x}) = P(T = t \mid T \geq t, \mathbf{x})$.

When working with a PW approach, see Holford (1980) and Laird and Olivier (1981), the follow-up is decomposed into J intervals. The hazard rate, defined in this case as $\lambda(t \mid \mathbf{x}) = \lim_{\Delta t \to 0} \frac{1}{\Delta t} P(t \leq T < t + \Delta t \mid T \geq t, \mathbf{x})$, is assumed to be constant within each interval. An illustration of the aforementioned discretised follow-up and hazard can be found in Figure 1 of Bender et al. (2018). In order to achieve correspondence with the Bernoulli or the Poisson log-likelihoods, we require an augmentation of the original dataset into a "long" format. See Bender et al. (2018) (Table 1 and Table 2) as well as Berger and Schmid (2018) (Equations (3.9) and (3.10)) for an illustration of the augmented data for PW and DT approaches, respectively. In practice, we transform the data using the `R` packages `discSurv` (Welchowski et al., 2022, DT) and `pammtools` (Bender and Scheipl, 2018, PW). This produces the following

auxiliary variable:

$$\delta_j = \begin{cases} 1, & \text{if subject has an event in } \mathcal{I}_j, \\ 0, & \text{else,} \end{cases} \quad \text{with} \quad \mathcal{I}_j = \begin{cases} [\kappa_{j-1},\ \kappa_j) & \text{if } DT, \\ (\kappa_{j-1},\ \kappa_j] & \text{if } PW, \end{cases}$$

Using the augmented data, the log-likelihood of a *DT* model coincides with the Bernoulli log-likelihood (Tutz and Schmid, 2016). The discrete hazard is then $\lambda(j \mid \mathbf{x}_j) = \vartheta_j = g^{-1}(\eta_j)$, where $\vartheta_j \in [0, 1]$ is the parameter of the Bernoulli distribution. If one follows a *PW* approach, the log-likelihood contribution coincides with the Poisson log-likelihood with offset $\ln(t_j) = o_j$, where t_j is the time spent in the j-th interval (Bender et al., 2018). In this case, the hazard is $\lambda(j \mid \mathbf{x}_j) = \exp(\eta_j) = \vartheta_j / \exp(o_j)$, where $\vartheta_j = \exp(\eta_j) \exp(o_j)$ is the parameter of the Poisson distribution. For either approach, a structured additive predictor is set as foundation for a regression model for the hazard rate:

$$\lambda(t \mid \mathbf{x}) = g^{-1}\Big(\underbrace{s_0(t)}_{\text{link-scale baseline hazard}} + \sum_{r=1}^{P} s_r(t, x_r) \Big) \quad \forall\ t \in\ \mathcal{I}_j,$$

where the baseline hazard is modelled as a smooth function of time and the functional form of the covariates can be any of those described in Section 2.1. For *PW* models $g(\cdot)$ is the natural logarithm function. In the *DT* framework, $g(\cdot)$ can be any suitable link function for for parameters with range $[0, 1]$, e.g. *logit*, *probit*, and *clog-log*. We remark that our implementation currently features only the clog-log link but we plan to add others in the future. From a practical perspective, the clog-log link produces the "grouped proportional hazards model", which can be seen as a discretised version of Cox's proportional hazards model (Tutz and Schmid, 2016).

3 Embedding DT and PW approaches into distributional copula regression

Using the *DT* or *PW* approach together with the augmented data, the i-th observation of the time-to-event margin is now represented by $j = 1, \ldots, j(i)$ auxiliary variables, where $j(i)$ denotes the length of the sequence emanating from the i-th observational unit in the sample. We denote these as $\boldsymbol{\delta} = \big(\delta_1, \ldots, \delta_{j(i)}\big)^\top$, which can have two possible expressions: $\boldsymbol{\delta} = \mathbf{0}_{j(i)}^\top$ in case of censoring or $\boldsymbol{\delta} = \big(\mathbf{0}_{j(i)-1},\ 1\big)^\top$ in case of an event, i.e. a non-censored observation, where $\mathbf{0}_{j(i)}^\top$ and $\mathbf{0}_{j(i)-1}^\top$ denote zero vectors of length $j(i)$ and $j(i)-1$, respectively. We propose to use the following function to model the hazard rate of the time-to-event variable using either

DT or *PW* techniques:

$$F\left(\boldsymbol{\delta}\right) = \begin{cases} f\left(\mathbf{0}_{j(i)}\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^{\top} \\ f\left(\mathbf{0}_{j(i)}\right) + f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1},\ 1\right)^{\top}, \end{cases} \tag{3}$$

where the expression of $f(\boldsymbol{\delta})$ depends on the adopted approach. For instance, choosing a *DT* model and setting the link function to the clog-log, i.e. $\lambda(j;\mathbf{x}_j) = \vartheta_j = 1 - \exp(-\exp(\eta_j))$, results in the following function based on the Bernoulli likelihood:

$$f\left(\boldsymbol{\delta}\right) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^{\top}, \\ \exp\left(-\sum_{j=1}^{j(i)-1} \exp(\eta_j)\right) - \exp\left(-\sum_{j=1}^{j(i)} \exp(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1},\ 1\right)^{\top}. \end{cases} \tag{4}$$

Using a *PW* approach, the conditional hazard corresponds to $\lambda(j;\mathbf{x}_j) = \exp(\eta_j) = \vartheta_j / \exp(o_j)$, leading to the following function based on the Poisson likelihood:

$$f(\boldsymbol{\delta}) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)} \vartheta_j\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}^{\top}, \\ \exp\left(-\sum_{j=1}^{j(i)} \vartheta_j\right) \vartheta_{j(i)}, & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1},\ 1\right)^{\top}. \end{cases} \tag{5}$$

We provide analytical first and second order partial derivatives of the proposed functions $f(\boldsymbol{\delta})$ and $F(\boldsymbol{\delta})$ based on *DT* and *PW* approaches w.r.t. a generic coefficient vector $\boldsymbol{\beta}$ in Supplement B. These are required for the simultaneous estimation algorithm of our software implementation, which is based entirely on `GJRM` (Marra and Radice, 2023).

3.1 Bivariate additive copula regression for mixed non-time-to-event & time-to-event responses

Recall that the bivariate non-commensurable outcome of interest is comprised of a non-time-to-event and a potentially right-censored time-to-event variable. We now replace the survival function of the potentially right-censored time-to-event margin with the proposed function $F(\boldsymbol{\delta})$ in Equation (1). Note that in this case Sklar's theorem is no longer used in its original form. We focus on cases when the non-time-to-event margin is binary, therefore the log-likelihood of the i-th observation has the following form (Marra et al., 2020):

$$\begin{aligned} \ell = (1-y) \ln \Big\{ C[F_1(0), F_2(\boldsymbol{\delta}); \vartheta^{(c)}] - C[F_1(0), F_2(\boldsymbol{\delta}) - f_2(\boldsymbol{\delta}); \vartheta^{(c)}] \Big\} + \\ y \ln \Big\{ f_2(\boldsymbol{\delta}) - C[F_1(0), F_2(\boldsymbol{\delta}); \vartheta^{(c)}] + C[F_1(0), F_2(\boldsymbol{\delta}) - f_2(\boldsymbol{\delta}); \vartheta^{(c)}] \Big\}, \end{aligned} \tag{6}$$

where the copula $C(\cdot)$ is evaluated using the CDF of y at zero, i.e. $F_1(0)$, and $F_2(\boldsymbol{\delta})$.

Considerations for copula modelling: A crucial aspect of copula modelling is the fact that the dependence structure between the margins is preserved if a monotonic transformation is applied to them (Nelsen, 2006; Klement et al., 2002). If the proposed function $F(\boldsymbol{\delta})$ is a monotonic decreasing function of the time-to-event variable T, then the original dependence structure will be preserved and the use of our proposed approach would be justified. Otherwise, $F(\boldsymbol{\delta})$ will alter the original dependence structure, rendering our proposal not suitable for copula modelling. Supplement A contains a proof that the proposed function $F(\boldsymbol{\delta})$ is a monotonic decreasing function of the original variable for both DT and PW approaches. Hence, replacing the survival function $S(t)$ with $F(\boldsymbol{\delta})$ in a bivariate copula model preserves the original dependence structure of interest. Additionally, we have conducted various simulation studies in order to assess the performance of the proposed approach using DT and PW likelihoods, these can be found in Supplement D.

4 Analysis of red-light running behaviour of E-cyclists

We study the behaviour of E-bike cyclists (E-cyclists) regarding traffic red-light running or traffic red light violation using a sample of $n = 2{,}173$ subjects from Shanghai, China collected and previously analysed by Gao et al. (2020). The margins of the bivariate mixed outcome consist of a binary non-time-to-event and a time-to-event variable, respectively. The binary response `VisualSearch` indicates whether a cyclist quickly turns their head in both directions while waiting at the crossing (0=no, 1=yes). This variable is considered a relevant factor in describing cyclists' red-light running and driving behaviour (Fraboni et al., 2018; Rupi and Krizek, 2019). The time-to-event variable `WaitingTime` gives the time in seconds until an E-cyclist crosses an intersection when the traffic light is red. The censoring indicator `RedLight` is equal to one if an E-cyclist runs the red-light, and zero if they wait until the traffic light turns green. This means that those E-cyclists that arrived at the intersection during a red-light and only cross it until the signal turns green are treated as censored observations (Gao et al., 2020). This yields a censoring rate of 47.78%. It could be argued that those individuals that do not run the red traffic light in the data, i.e. censored observations, will never commit such an infringement and should not be analysed jointly with those that do so. However, we believe it is more reasonable to assume that any E-cyclist is capable of running a red traffic light *under certain circumstances*. Hence, for the censored observations in the sample, the conditions for violating the red traffic light were simply not met at the time of collecting the data and we assume that under different conditions such cyclists would cross an intersection given a red-light. The covariates in the data describe three aspects of the E-cyclists: Individual characteristics, social influence and cycling information (Gao et al., 2020). Among the individual characteristics are the `Gender` (female=0, male=1) of the cyclist and the factor

Table 1: Description of the variables in the data. Summary column shows median survival time for `WaitingTime` and percentages of the remaining variables.

Variable	Description	Summary
`WaitingTime`	Time (in seconds) waiting at crossing for green traffic light.	79
`RedLight`	Red-light running (0 = no, 1 = yes).	47.8/52.2
`VisualSearch`	Cyclist turns head quickly in both directions (0=no, 1=yes).	41.1/58.9
`Gender`	0=Female, 1=Male.	29.8/70.2
`RiderType`	Type of cyclist riding the e-bike (0=Normal, 1=Takeaway).	51.9/48.1
`GroupSize`	Number of cyclists waiting at intersection at arrival of cyclist. ($\in$ {"none", "1 to 4", "5 and more"}).	33.5/55.4/11.1
`ConformBehaviour`	Another cyclist runs the red light (0=no, 1=yes).	50.3/49.7
`Position`	Position at crossing where cyclist waits. ($\in$ {"behind stop line", "middle", "close to motorised lane"}).	29.5/34.4/36.1
`ComingDirection`	Direction from where the cyclist comes into intersection. ($\in$ {"through", "left", "right"}).	63.6/20.6/15.7

Sample size: n = 2,173. Censoring rate: 47.78%.

variable `RiderType` which indicates whether the cyclist is a takeaway delivery driver or not. The social influence aspect is represented by two variables.

The categorical variable `GroupSize`, which says how many other cyclists were waiting at the intersection at the arrival of the subject ("none", "1 to 4" and "five and more"), and `ConformBehaviour` which is equal to one if another cyclist runs the red light and zero otherwise. Social influence characteristics have been found to play an important role in decreasing the propensity of red-light running (Bai and Sze, 2020). Lastly, the cycling information is described by the categorical variables `Position` and `ComingDirection`. Previous studies have found that `Position` plays an important role in determining driving behaviour (Tang et al., 2020). See Table 1 for a description of the variables.

Model configuration: We aim to model the joint distribution of `VisualSearch` and `WaitingTime` using our proposed distributional copula regression approach. The binary response `VisualSearch` is modelled using a Bernoulli distribution with parameter $\vartheta^{(1)} \in [0, 1]$. The hazard rate of the time-to-event variable `WaitingTime` is modelled using the proposed functions based on the auxiliary variables $\boldsymbol{\delta}$ using DT and PW approaches:

$$(\texttt{VisualSearch},\ \texttt{WaitingTime})^\top \sim C[F_1(\texttt{VisualSearch};\ \vartheta^{(1)}),\ F_2(\boldsymbol{\delta};\ \boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}],$$

where $F_1(\cdot)$ is the CDF of the Bernoulli distribution and $F_2(\cdot)$ is the proposed function from Equation (3) with $\boldsymbol{\vartheta}^{(2)}$ denoting the hazard rate built using either DT (clog-log link) or PW (natural logarithm link) methods. We omit the index k in the distribution parameters for the remainder of the manuscript due to both margins depending only on one parameter. In order to construct the model of the hazard for `WaitingTime`, the follow-up is split using 20 intervals, resulting in intervals that are 6.5 seconds of length. We investigated alternative configurations with 130, 65, 35 and 13 intervals, but these did not change the results in a

Table 2: Akaike and Bayesian information criteria of copula models with clog-log link for `VisualSearch` as well as independent margins. Bold numbers indicate the best-fitting model.

	DT		*PW*	
Copula	AIC	BIC	AIC	BIC
Independence	9266.31	9479.82	19544.90	19730.12
Gaussian	9159.60	9342.00	19825.64	20014.59
Frank	9157.01	9339.28	19442.84	19625.98
Clayton	**9155.67**	**9338.02**	**19441.87**	**19625.11**
Clayton (90°)	9272.28	9454.43	19550.75	19733.74
Gumbel	9166.21	9348.63	19491.04	19674.13
Gumbel (90°)	9272.12	9454.26	19548.36	19731.35

relevant way. The structured additive predictors of the model are:

$$\begin{aligned}
\eta^{(1)} &= \beta_0^{(1)} + \beta_1^{(1)}\texttt{Gender} + \beta_2^{(1)}\texttt{RiderType} + \beta_3^{(1)}\texttt{Position},\\
\eta^{(2)} &= s_0^{(2)}(t,\ \texttt{RiderType}) + \beta_1^{(2)}\texttt{Gender} + \beta_2^{(2)}\texttt{RiderType} + \beta_3^{(2)}\texttt{Position} + \beta_4^{(2)}\texttt{GroupSize} +\\
&\quad \beta_5^{(2)}\texttt{ConformBehaviour},\\
\eta^{(c)} &= \beta_0^{(c)} + \beta_1^{(c)}\texttt{RiderType} + \beta_2^{(c)}\texttt{ConformBehaviour},
\end{aligned}$$

where the link-scale baseline hazard $s_0^{(2)}(\cdot)$ is modelled as a smooth function of time using P-splines with second order differences and is stratified depending on `RiderType`. Hence, we fit a separate baseline hazard for `Normal` and `TakeAway` E-cyclists. The remaining categorical covariates enter the model as linear effects with dummy coding. The dependence between the margins modelled via $\eta^{(c)}$ is allowed to change based on `RiderType` and `ConformBehaviour`.

The variable `ComingDirection` is not included in the model due having no significant effect on any additive predictor. Note that the covariates were selected into the different additive predictors based on hypothesis testing (Marra and Radice, 2017). The copula function is selected out of the Gaussian, Frank, Clayton and Gumbel (with 90° rotations) copulas by means of the AIC and BIC, see Table 2. We remark that `GJRM` supports more copula functions aside from the aforementioned, however we do not consider the entire catalogue in this manuscript. All model coefficients are estimated simultaneously. Confidence intervals are obtained as described in Radice et al. (2016). The goodness-of-fit of $\boldsymbol{\delta}$ is assessed using Cox-Snell residuals (Klein and Moeschberger, 2003) in Figure 2. An additional diagnostic of the fitted copula function can be found in Supplement C2.

Results using DT approach: The best fitting model is a Clayton copula with clog-log link for `VisualSearch`, the estimated coefficients are shown in the *DT* column of Table 3. The column corresponding to the *DT* models in Table 2 shows that the Clayton copula fits the data better compared to using independent margins. We find that on average there exists a significant "moderate" positive dependence between `VisualSearch` and `WaitingTime` with

Table 3: Estimated coefficients with 95% confidence intervals of the Clayton copula models using *DT* and *PW* approaches. All coefficients are estimated simultaneously.

	DT		*PW*	
Binary margin `VisualSearch`				
$\beta_0^{(1)}$	−0.675	[−0.827; −0.528]	−0.674	[−0.819; −0.529]
`GenderMale`	0.162	[0.023; 0.304]	0.164	[0.017; 0.309]
`RiderTypeTakeAway`	0.510	[0.386; 0.634]	0.504	[0.376; 0.628]
`PositionMiddle`	0.256	[0.109; 0.404]	0.257	[0.110; 0.400]
`PositionCloseMotorLane`	0.291	[0.146; 0.436]	0.297	[0.151; 0.442]
Time-to-event margin `WaitingTime`				
`GenderMale`	0.300	[0.134; 0.464]	0.290	[0.125; 0.454]
`RiderTypeTakeAway`	1.460	[1.294; 1.626]	1.496	[1.329; 1.663]
`PositionMiddle`	0.428	[0.258; 0.601]	0.420	[0.251; 0.589]
`PositionCloseMotorLane`	0.688	[0.525; 0.853]	0.683	[0.521; 0.845]
`GroupSize1to4`	−0.332	[−0.452; −0.215]	−0.356	[−0.476; −0.237]
`GroupSize5AndMore`	−0.758	[−0.996; −0.513]	−0.811	[−1.049; −0.580]
`ConformBehaviourYes`	0.276	[0.156; 0.397]	0.259	[0.136; 0.381]
Dependence				
$\beta_0^{(c)}$	0.546	[0.222; 0.873]	0.495	[0.160; 0.824]
`RiderTypeTakeAway`	−2.137	[−2.950; −1.315]	−2.560	[−3.576; −1.547]
`ConformBehaviourYes`	−0.711	[−1.244; −0.182]	−0.719	[−1.261; −0.160]
Kendall's $\hat{\tau}$	0.236	[0.181; 0.311]	0.217	[0.162; 0.292]

Sample size: $n = 2{,}173$. Censoring rate: 47.78%.

Kendall's τ estimated at $\hat{\tau} = 0.236$ [0.181, 0.311] with $\alpha = 0.05$. The additive predictor of the dependence parameter $\eta^{(c)}$ decreases for takeaway relative to normal cyclists. A less pronounced decrease in the predictor $\eta^{(c)}$ is also observed when another cyclist at the intersection commits a red traffic light violation (`ConformBehaviour` = 1). These estimates indicate that the behaviour of E-cyclists is influenced by social or group factors while waiting at an intersection. Table 4 shows the values of the estimated Kendall's τ for the combinations of the binary variables `RiderType` and `ConformBehaviour`. It can be seen that the dependence between the margins decreases for takeaway cyclists whenever there is conforming behaviour. Comparatively, the estimated dependence between `VisualSearch` and `ConformBehaviour` is much stronger for normal cyclists and when no other driver crosses the junction. Regarding the model for `VisualSearch`, takeaway cyclists have on average a significantly higher probability of checking for traffic in both directions relative to normal cyclists, *ceteris paribus*. On average, male cyclists also have a significantly higher change of conducting `VisualSearch` relative to female cyclists, *ceteris paribus*. Figure 1 shows the estimated hazard rate and corresponding estimated survival function from the Clayton model based on different values of `GroupSize` and `RiderType`. Overall, a noticeable shift upwards in the hazard rate can be

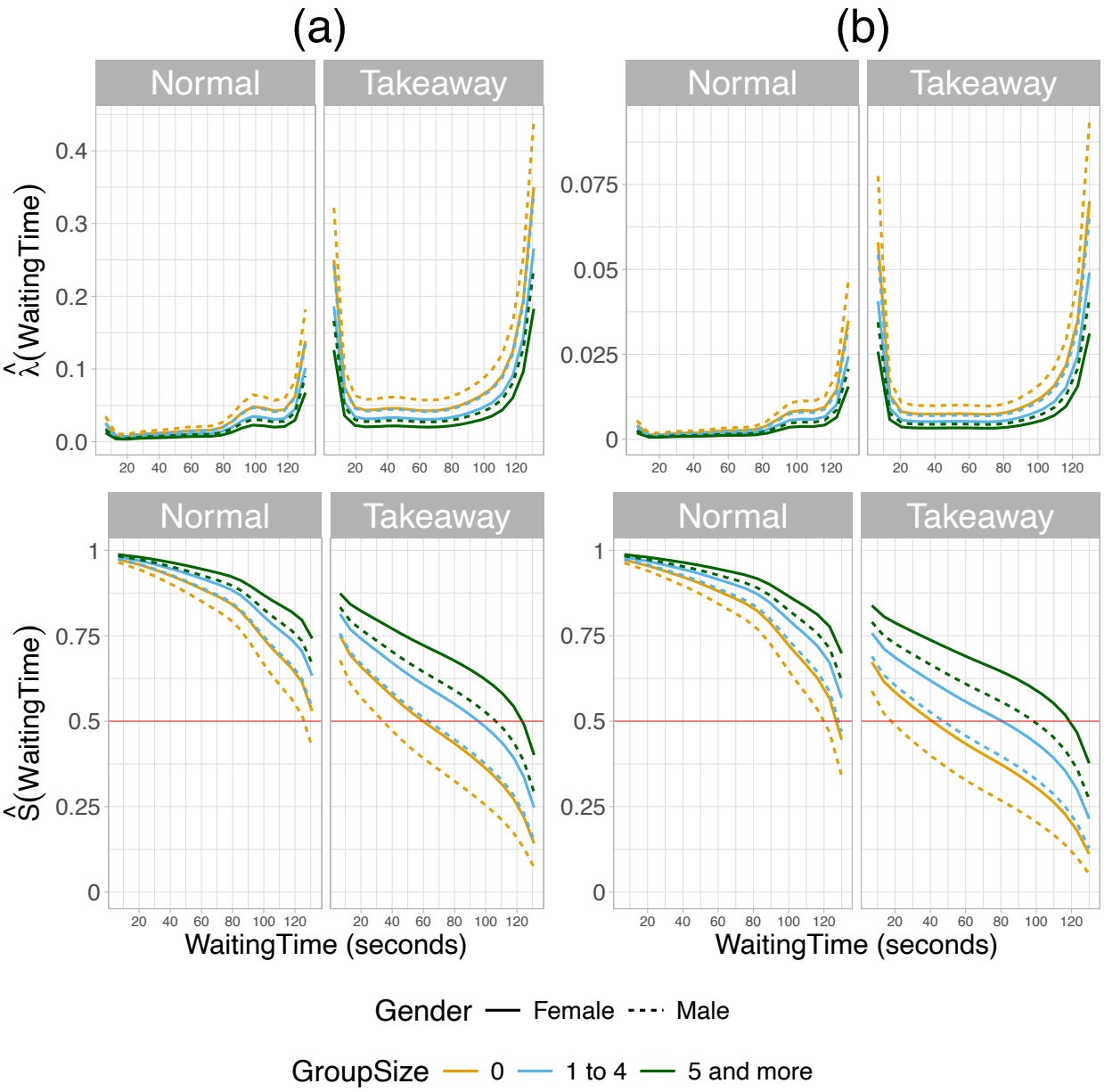

Figure 1: Estimated hazard rate and survival function across `RiderType`, `GroupSize` and `Gender` for *DT* (a) and *PW* (b) approaches. Red line indicates the median.

seen for takeaway cyclists. Conversely, the estimated survival function for takeaway cyclists lies below the one corresponding to normal cyclists. This phenomenon depicts the riskier behaviour that delivery-service cyclists exhibit in order to meet demand in the delivery-service industry (Gao et al., 2020; Zhang and Liu, 2023).

The `GroupSize` covariate describes the social behaviour of the cyclist while waiting for green traffic light. The magnitude of the estimated effect on the link-scale hazard rate increases based on the level of `GroupSize`. Hence, the more individuals waiting at the crossing or the more crowded the crossing becomes, the less likely that the cyclist will violate the traffic red-light. Figure 1 column (a) depicts the estimated hazard rates and survival functions based

Table 4: Estimated dependence expressed as Kendall's τ for the combination of `RiderType` and `ConformBehaviour` obtained using DT and PW approaches with 95% confidence intervals shown underneath the estimates.

	DT		PW	
`RiderType`	`Normal`	`TakeAway`	`Normal`	`TakeAway`
`ConformBehaviour`				
`No`	0.463	0.092	0.451	0.060
	[0.382; 0.545]	[0.042; 0.186]	[0.370; 0.532]	[0.022; 0.149]
`Yes`	0.298	0.048	0.286	0.030
	[0.213; 0.402]	[0.023; 0.097]	[0.203; 0.386]	[0.012; 0.074]

on the DT approach. The effect of `GroupSize` can be seen to produce a shift downwards in the estimated hazards and upwards in the estimated survival functions. The estimated survival functions shown in Figure 1 column (a) exhibit different patters for normal and takeaway cyclists, respectively. It can be seen that the median survival time is dramatically lower for takeaway cyclists. The estimated hazard rate shown in Figure 1 column (a) shows that the hazard for takeaway E-cyclists has a "U" or bathtub shape. This means that the risk of violating the red traffic light is higher at the cyclist's arrival at the intersection as well as close to the end of the red signal. The Cox-Snell residuals shown in Figure 2 (a) corresponding to the DT approach indicate that the fitted model with a stratified baseline hazard for `RiderType` fits the data well. Table C1 and Figure C1 in Supplement C1 show the estimated coefficients as well as the estimated hazard and survival functions, respectively, when assuming independence between the margins. Overall, the estimated coefficients from univariate independent models exhibit some differences in their magnitude, but a similar trend can be seen for all of the considered covariates. The diagnostics regarding the fitted Clayton copula function in Supplement C2 indicate that the Clayton copula is adequate to model the data. However, we remark that some asymmetry in the margins can be seen in some instances.

Results using PW approach: According to Table 2, the best-fitting copula is once again the Clayton copula. In addition, the column PW in Table 2 shows that the Clayton copula model delivers a better fit compared to fitting independent univariate margins. The estimated coefficients for the model based on the PW approach closely resemble those of the DT approach. The 95% confidence intervals shown in Table 3 are also very similar for both approaches. The estimated dependence quantified as Kendall's τ is also close to the DT approach-based model, sitting at $\hat{\tau} = 0.217 \in [0.162; 0.292]$. See Table 4 for the estimated values of Kendall's τ for every combination of `RiderType` and `ConformBehaviour`. Once again, the estimated values resemble that of the model based on the DT approach. The estimated hazard rates and survival functions shown in Figure 1(b) exhibit the same pattern as those based on the DT approach. The Cox-Snell residuals from Figure 2(b) also suggest that the PW-based model fits the data well apart from some outliers in the upper tail of the data. The results from fitting independent univariate models shown in Table C1 and Figure C1 in

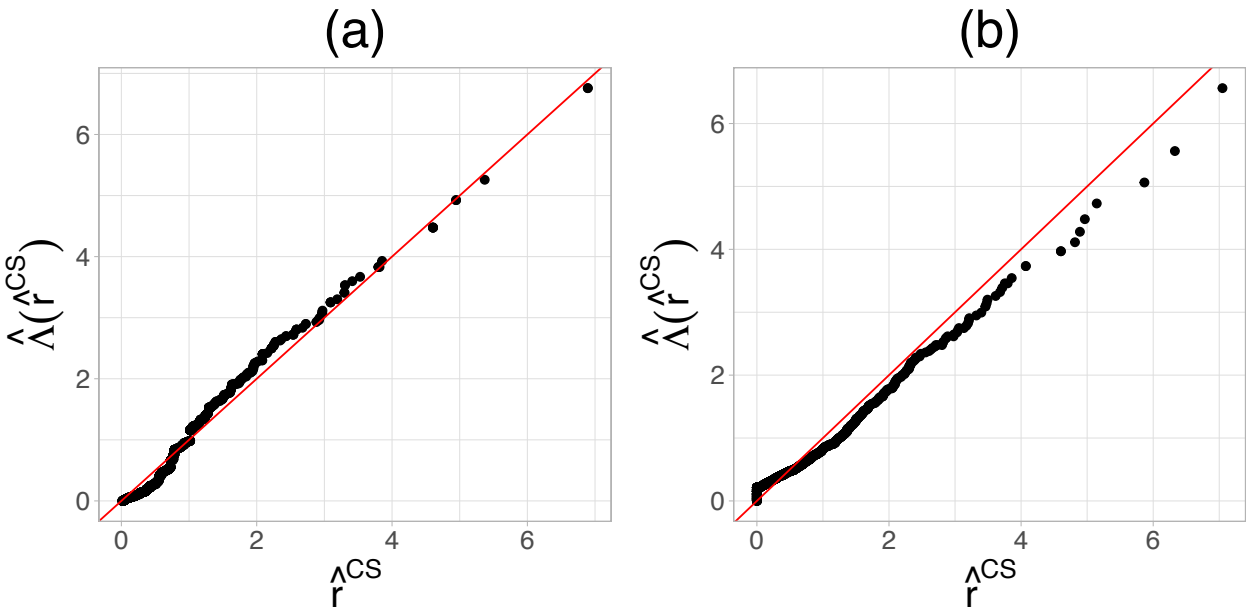

Figure 2: Cox-Snell residuals corresponding to the model of `WaitingTime` using the *DT* (a) and *PW* (b) approaches.

Supplement C1 exhibit some changes in the magnitude of the estimated coefficients. Similar trends to those found in the copula models can be observed. Once again, the diagnostics regarding the Clayton copula from Supplement C2 point towards an adequate fit. Similar to the model using the *DT* approach, some asymmetry can be seen in the margins. One possible solution to accommodate the aforementioned asymmetry would be to implement so-called "Khoudraji copulas" (Genest et al., 1998), which modify copula functions by introducing two scalar asymmetry parameters. However, this is outside of the scope of this manuscript.

5 Discussion

We have proposed a distributional regression model that embeds discrete time-to-event as well as piecewise-exponential modelling approaches into a copula regression framework for bivariate mixed outcomes. This allows to cast highly flexible regression models for the hazard rate of the time-to-event margin, which can include smooth functions of time, time-dependent covariates as well as time-varying effects. From a computational point of view, since our regression model for the time-to-event margin is based on the hazard rate, it does not require any constraints on the estimated smooth functions of time. Hence, we can avoid adding penalties or reparameterisations to impose such constraints, e.g. baseline function of time must be monotonically decreasing, see Liu et al. (2018) and Marra and Radice (2019) for more details. The data transformation into long format depends on the configurations set by the analyst. Instead of treating the number of intervals as a tuning parameter, we suggest to set the number of intervals that cut the follow-up to a rather large value (e.g. 20 intervals) and let the penalty term of the baseline hazard take care of the compromise between overfitting the data

and smoothness of the estimated function.

We embedded the *DT* and *PW* approaches into a bivariate distributional copula regression framework and implemented them in the convenient open source `R` library `GJRM`, allowing practitioners to cast flexible statistical models for mixed outcomes as well as bivariate time-to-event (implemented, but not presented here). Although not shown in this manuscript, the proposed approach can also be used for modelling bivariate right-censored time-to-event responses as in Marra and Radice (2019), using *DT* or *PW* methods. This would involve modifying the log-likelihood to one based on bivariate discrete data shown in van der Wurp et al. (2020). We have demonstrated the versatility of the proposed approach by analysing the joint bivariate distribution of a non-commensurable outcome that consisted of binary and time-to-event margins using *DT* and *PW* approaches. The analysis showed that takeaway E-cyclists have a higher propensity to conduct a visual search, i.e. inspect both directions of an intersection, as well as to run traffic red lights compared to normal E-cyclists. The model shows a moderately strong estimated dependence between visual search and the time to red-light running for all E-cyclists. Our results also show that the red-light running of another cyclist waiting at the same intersection decreases the dependence between the responses.

Our approach has several potential avenues for future developments. One of them would be to adapt our software implementation to support other non-time-to-event margins such as ordinal outcomes following Hohberg et al. (2021). Another aspect could be the inclusion of a *cure fraction* (i.e. *cure models*, Peng and Yu, 2021) to account for subjects that will never experience the event. This phenomenon appears to be present in the data analysed in Section 4, where cyclists that did not violate the traffic light are considered censored observations. Instead, one could argue that those cyclists belong to a sub-population that does not commit such infractions. An indication of the presence of such a cure fraction can be seen in the fact that the estimated survival function of normal (non-takeaway) E-cyclists remains far from zero. Thus, there may be E-cyclists that "never" violate red traffic lights. We may consider relaxing the assumption of non-informative censoring times and model their distribution similar to Dettoni et al. (2020). Alternatively, we could explore relaxing the assumption of independent censoring and allow for dependent censoring as in Czado and Van Keilegom (2022). By doing so we would need to not only account for the dependence between the margins but also between the censoring and event times, which would result in a model that uses two copula functions. Another potential extension is modelling recurrent events based on Ramjith et al. (2022), which could require modifications of the proposed *PW* functions, but it would greatly extend the applicability of our methods. Data-driven variable selection could be tackled, for example, using a quadratic approximation of LASSO-type penalised models as in van der Wurp and Groll (2021), or boosting for distributional copula regression following Hans et al. (2022) and Strömer et al. (2023). Lastly, more flexible types of copula functions could be explored. For instance, "Khoudraji copulas" (Genest et al., 1998),

which allow to model asymmetry in the margins. These type of copulas are an attractive candidate for future work due to their parametric nature.

Acknowledgements

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) as project 271512359.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Bai, L. and Sze, N. (2020). Red light running behavior of bicyclists in urban area: Effects of bicycle type and bicycle group size. *Travel Behaviour and Society*, 21:226–234.
URL: https://doi.org/10.1016/j.tbs.2020.07.003.

Bender, A., Groll, A., and Scheipl, F. (2018). A generalized additive model approach to time-to-event analysis. *Statistical Modelling*, 18(3-4):299–321.
URL: https://doi.org/10.1177/1471082X17748083.

Bender, A. and Scheipl, F. (2018). *pammtools: Piece-wise exponential Additive Mixed Modeling tools*. R package version 0.5.93.
URL https://CRAN.R-project.org/package=pammtools.

Berger, M. and Schmid, M. (2018). Semiparametric regression for discrete time-to-event data. *Statistical Modelling*, 18(3-4):322–345.
URL: https://doi.org/10.1177/1471082X17748084.

Czado, C. and Van Keilegom, I. (2022). Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738.
URL: https://doi.org/10.1093/biomet/asac067.

de Leon, A. R. and Wu, B. (2011). Copula-based regression models for a bivariate mixed discrete and continuous outcome. *Statistics in Medicine*, 30(2):175–185.
URL: https://doi.org/10.1002/sim.4087.

Dettoni, R., Marra, G., and Radice, R. (2020). Generalized link-based additive survival models with informative censoring. *Journal of Computational and Graphical Statistics*, 29(3):503–512.
URL: https://doi.org/10.1080/10618600.2020.1724544.

Eilers, P. H. C. and Marx, B. D. (1996). Flexible smoothing with B-splines and penalties. *Statistical Science*, 11(2):89–121.
URL: https://doi.org/10.1214/ss/1038425655.

Fraboni, F., Puchades, V. M., Angelis, M. D., Pietrantoni, L., and Prati, G. (2018). Red-light running behavior of cyclists in Italy: An observational study. *Accident Analysis & Prevention*, 120:219–232.
URL: https://doi.org/10.1016/j.aap.2018.08.013.

Gao, X., Zhao, J., and Gao, H. (2020). Red-light running behavior of delivery-service e-cyclists based on survival analysis. *Traffic Injury Prevention*, 21(8):558–562.
URL: https://doi.org/10.1080/15389588.2020.1819989.

Genest, C., Ghoudi, K., and Rivest, L.-P. (1998). "Understanding Relationships Using Copulas", by Edward Frees and Emiliano Valdez. *North American Actuarial Journal*, 2(3):143–149.
URL: https://doi.org/10.1080/10920277.1998.10595749.

Groll, A., Kneib, T., Mayr, A., and Schauberger, G. (2018). On the dependency of soccer scores – A sparse bivariate poisson model for the UEFA european football championship 2016. *Journal of Quantitative Analysis in Sports*, 14(2):65–79.
URL: https://doi.org/10.1515/jqas-2017-0067.

Hans, N., Klein, N., Faschingbauer, F., Schneider, M., and Mayr, A. (2022). Boosting distributional copula regression. *Biometrics*, 79:2298–2310.
URL: https://doi.org/10.1111/biom.13765.

Hohberg, M., Donat, F., Marra, G., and Kneib, T. (2021). Beyond unidimensional poverty analysis using distributional copula models for mixed ordered-continuous outcomes. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 70(5):1365–1390.
URL: https://doi.org/10.1111/rssc.12517.

Holford, T. R. (1980). The analysis of rates and of survivorship using log-linear models. *Biometrics*, 36(2):299–305.
URL: https://doi.org/10.2307/2529982.

Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data by using bivariate poisson models. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 52(3):381–393.
URL: https://doi.org/10.1111/1467-9884.00366.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis*. Springer New York.
URL: https://doi.org/10.1007/b97377.

Klein, N., Kneib, T., Klasen, S., and Lang, S. (2014). Bayesian structured additive distributional regression for multivariate responses. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 64(4):569–591.
URL: https://doi.org/10.1111/rssc.12090.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S., and McGovern, M. E. (2018). Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, 38(3):413–436.
URL: https://doi.org/10.1002/sim.7985.

Klement, E. P., Mesiar, R., and Pap, E. (2002). Invariant copulas. *Kybernetika*, 38(3):275–285.
URL: http://eudml.org/doc/33582.

Kocherlakota, S. and Kocherlakota, K. (2017). *Bivariate Discrete Distributions*. CRC Press.
URL: https://doi.org/10.1201/9781315138480.

Lai, C. D. and Balakrishnan, N. (2009). *Continuous Bivariate Distributions*. Springer New York.
URL: https://doi.org/10.1007/b101765.

Laird, N. and Olivier, D. (1981). Covariance analysis of censored survival data using log-linear analysis techniques. *Journal of the American Statistical Association*, 76(374):231–240.
URL: https://doi.org/10.2307/2287816.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546.
URL: https://doi.org/10.1177/0962280216664760.

Marra, G. and Radice, R. (2017). Bivariate copula additive models for location, scale and shape. *Computational Statistics and Data Analysis*, 112:99–113.
URL: https://doi.org/10.1016/j.csda.2017.03.004.

Marra, G. and Radice, R. (2019). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115(530):886–895.
URL: https://doi.org/10.1080/01621459.2019.1593178.

Marra, G. and Radice, R. (2023). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-5.
URL https://CRAN.R-project.org/package=GJRM.

Marra, G., Radice, R., and Zimmer, D. M. (2020). Estimating the binary endogenous effect of insurance on doctor visits by copula-based regression additive models. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 69(4):953–971.
URL: https://doi.org/10.1111/rssc.12419.

Medina-Olivares, V., Calabrese, R., Crook, J., and Lindgren, F. (2023). Joint models for longitudinal and discrete survival data in credit scoring. *European Journal of Operational Research*, 307(3):1457–1473.
URL: https://doi.org/10.1016/j.ejor.2022.10.022.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2022). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.5.0.
URL: https://CRAN.R-project.org/package=VineCopula.

Najib, M. K., Nurdiati, S., and Sopaheluwakan, A. (2022). Multivariate fire risk models using copula regression in Kalimantan, Indonesia. *Natural Hazards*, 113(2):1263–1283.
URL: https://doi.org/10.1007/s11069-022-05346-3.

Nelsen, R. B. (2006). *An Introduction to Copulas*. Springer New York.
URL: https://doi.org/10.1007/0-387-28678-0.

Peng, Y. and Yu, B. (2021). *Cure Models*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/9780429032301.

Radice, R., Marra, G., and Wojtyś, M. (2016). Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5):981–995.
URL: http://dx.doi.org/10.1007/s11222-015-9581-6.

Ramjith, J., Bender, A., Roes, K. C. B., and Jonker, M. A. (2022). Recurrent events analysis with piece-wise exponential additive mixed models. *Statistical Modelling*.
URL: https://doi.org/10.1177/1471082X221117612.

Rappl, A., Kneib, T., Lang, S., and Bergherr, E. (2023). Spatial joint models through Bayesian structured piecewise additive joint modelling for longitudinal and time-to-event data. *Statistics and Computing*, 33(6):135.
URL: https://doi.org/10.1007/s11222-023-10293-5.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/b12208.

Rupi, F. and Krizek, K. J. (2019). Visual eye gaze while cycling: Analyzing eye tracking at signalized intersections in urban conditions. *Sustainability*, 11(21):6089.
URL: https://doi.org/10.3390/su11216089.

Spissu, E., Pinjari, A. R., Pendyala, R. M., and Bhat, C. R. (2009). A copula-based joint multinomial discrete–continuous model of vehicle type choice and miles of travel. *Transportation*, 36(4):403–422.
URL: https://doi.org/10.1007/s11116-009-9208-x.

Stasinopoulos, M. D., Rigby, R. A., Heller, G. Z., Voudouris, V., and Bastiani, F. D. (2017). *Flexible Regression and Smoothing*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/b21973.

Strömer, A., Klein, N., Staerk, C., Klinkhammer, H., and Mayr, A. (2023). Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*, 42(11):1779–1801.
URL: https://doi.org/10.1002/sim.9699.

Suresh, K., Taylor, J. M., and Tsodikov, A. (2021). A copula-based approach for dynamic prediction of survival with a binary time-dependent covariate. *Statistics in Medicine*, 40(23):4931–4946.
URL: https://doi.org/10.1002/sim.9102.

Suresh, K., Taylor, J. M. G., and Tsodikov, A. (2019). A gaussian copula approach for dynamic prediction of survival with a longitudinal biomarker. *Biostatistics*, 22(3):504–521.
URL: https://doi.org/10.1093/biostatistics/kxz049.

Tang, T., Wang, H., Ma, J., and Zhou, X. (2020). Analysis of crossing behavior and violations of electric bikers at signalized intersections. *Journal of Advanced Transportation*, 2020:1–14.
URL: https://doi.org/10.1155/2020/3594963.

Tutz, G. and Schmid, M. (2016). *Modeling Discrete Time-to-Event Data*. Springer International Publishing.
URL: https://doi.org/10.1007/978-3-319-28158-2.

Umlauf, N. and Kneib, T. (2018). A primer on Bayesian distributional regression. *Statistical Modelling*, 18(3-4):219–247.
URL: https://doi.org/10.1177/1471082X18759140.

van der Wurp, H. and Groll, A. (2021). Introducing LASSO-type penalisation to generalised joint regression modelling for count data. *AStA Advances in Statistical Analysis*, 107(1-2):127–151.
URL: https://doi.org/10.1007/s10182-021-00425-5.

van der Wurp, H., Groll, A., Kneib, T., Marra, G., and Radice, R. (2020). Generalised joint regression for count data: a penalty extension for competitive settings. *Statistics and Computing*, 30(5):1419–1432.
URL: https://doi.org/10.1007/s11222-020-09953-7.

Wagner, H. and Tüchler, R. (2013). Sparse bayesian modeling of mixed econometric data using data augmentation. In de Leon, A. R. and Chough, K. C., editors, *Analysis of Mixed Data*, pages 173–188. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/b14571.

Welchowski, T., Berger, M., Koehler, D., and Schmid, M. (2022). *discSurv: Discrete Time Survival Analysis*. R package version 2.0.0.
URL: https://CRAN.R-project.org/package=discSurv.

Wood, S. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/9781315370279.

Zhang, Z. and Liu, C. (2023). Exploration of riding behaviors of food delivery riders: A naturalistic cycling study in Changsha, China. *Sustainability*, 15(23).
URL: https://doi.org/10.3390/su152316227.

Supplementary Material

for

"Bivariate distributional copula regression for mixed non-time-to-event & time-to-event responses"

Contents

Part A: Proof that $F(\boldsymbol{\delta})$ is a monotonic decreasing function of time.

In this supplement, we prove that the proposed function $F(\boldsymbol{\delta})$ is a monotone decreasing function of the original time-to-event variable T, where $\boldsymbol{\delta}$ is a sequence of auxiliary variables created using either discrete time-to-event (DT) or piecewise-exponential methods (PW). The index $i = 1, \ldots, n$ denoting the observational units is suppressed to avoid clutter. We are interested in modelling the time-to-event variable $T \in \mathbb{R}_+$, which can be either in continuous or discrete time, i.e. $T \in \mathbb{R}$, $T \in \mathbb{N}$ or some finite subset of the positive integers. The time-to-event variable is subject to random, independent and non-informative right-censoring. We observe $\tilde{T} = \min\{T,\ T^{cens}\}$ and $\delta = \mathbb{1}\{T \leq T^{cens}\}$, where T is the true event time and T^{cens} is a censoring time. As mentioned in Section 2.2, the data-augmentation produces an auxiliary variable that together with the transformed data allows to use either the Bernoulli (DT) or Poisson (PW) likelihood in order to fit the model. The auxiliary variables are obtained by splitting the follow-up time into a finite amount of intervals $\mathcal{I}_j$ which may be equidistant or of distinct finite length:

$$\delta_j = \begin{cases} 1, & \text{if subject } i \text{ has an event in } \mathcal{I}_j, \\ 0, & \text{else,} \end{cases}, \quad \text{and} \quad \mathcal{I}_j = \begin{cases} [\kappa_{j-1},\ \kappa_j), & \text{if } DT, \\ (\kappa_{j-1},\ \kappa_j], & \text{if } PW. \end{cases}$$

This results in $\boldsymbol{\delta}$, which is a vector of length $j(i)$ that contains the auxiliary variables, where $j(i)$ denotes the interval in which observation i experienced a censoring or an event (final interval), i.e. $\delta = 0$ or $\delta = 1$. The vector $\boldsymbol{\delta}$ has the following two realisations:

$$\boldsymbol{\delta} = \begin{cases} \left(\mathbf{0}_{j(i)}\right)^\top, & \text{in case of a censored observation,} \\ \left(\mathbf{0}_{j(i)-1},\ 1\right)^\top, & \text{in case of an event,} \end{cases}$$

where $\mathbf{0}_{j(i)}$ denotes a zero vector of length $j(i)$. In the following, we provide a general expression for the functions $f(\boldsymbol{\delta})$ and $F(\boldsymbol{\delta})$. Then we will prove that the latter is a monotonically decreasing function of the original time-to-event variable. Given the possible expressions as well as the conditional independence of the components of $\boldsymbol{\delta}$, the function $f(\boldsymbol{\delta})$ is constructed using the likelihoods in the following fashion:

$$f(\boldsymbol{\delta}) = \prod_{j=1}^{j(i)} f(\delta_j \mid \vartheta_j),$$

where $f(\delta_j \mid \vartheta_j)$ are univariate Bernoulli or Poisson likelihoods. Thus for censored observations, i.e. $\delta = \delta_{j(i)} = 0$ we get:

$$f(\mathbf{0}_{j(i)}) = \prod_{j=1}^{j(i)} f(0 \mid \vartheta_j).$$

For non-censored observations, i.e. $\delta = \delta_{j(i)} = 1$ we obtain:

$$f\left((\mathbf{0}_{j(i)-1},\ 1)\right) = \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)}).$$

The proposed function $F(\boldsymbol{\delta})$ is obtained by aggregating over $\breve{\boldsymbol{\delta}} = \{\mathbf{0}_{j(i)}, (\mathbf{0}_{j(i)}, 1)\}$:

$$\begin{aligned} F(\mathbf{0}_{j(i)}) &= f(\mathbf{0}_{j(i)}), \\ F((\mathbf{0}_{j(i)-1},\ 1)) &= f(\mathbf{0}_{j(i)}) + f((\mathbf{0}_{j(i)-1},\ 1)). \end{aligned}$$

We will now prove that for both DT and PW approaches, the proposed function $F(\cdot)$ is a monotonically decreasing function of time, in other words: For $t < t^*$, it follows that $F(\boldsymbol{\delta}) \geq F(\boldsymbol{\delta}^*)$, where $\boldsymbol{\delta}$ is the vector of auxiliary variables emanating from t, and $\boldsymbol{\delta}^*$ corresponds t^*. There are four cases to be considered:

1. $\delta = 0$ (censored observation) and $t, t^* \in (\kappa_{j-1},\ \kappa_j]$ (PW) or $t, t^* \in [\kappa_{j-1},\ \kappa_j)$ (DT).
2. $\delta = 0$ (censored observation) and $t \in (\kappa_{j-1},\ \kappa_j]$, $t^* \in (\kappa_{j+h},\ \kappa_{j+h+1}]$ (PW) or $t \in [\kappa_{j-1},\ \kappa_j)$, $t^* \in [\kappa_{j+h},\ \kappa_{j+h+1})$ (DT), with $h = 0, 1, 2, 3, \ldots$.
3. $\delta = 1$ (non-censored observation) and $t, t^* \in (\kappa_{j-1},\ \kappa_j]$ (PW) or $t, t^* \in [\kappa_{j-1},\ \kappa_j)$ (DT).
4. $\delta = 1$ (non-censored observation) and $t \in (\kappa_{j-1},\ \kappa_j]$, $t^* \in (\kappa_{j+h},\ \kappa_{j+h+1}]$ (PW) or $t \in [\kappa_{j-1},\ \kappa_j)$, $t^* \in [\kappa_{j+h},\ \kappa_{j+h+1})$ (DT), with $h = 0, 1, 2, 3, \ldots$.

Case 1: $\delta = 0$ and $t, t^* \in (\kappa_{j-1},\ \kappa_j]$ (PW) or $t, t^* \in [\kappa_{j-1},\ \kappa_j)$ (DT). This implies that both vectors of auxiliary variables have the same length, i.e. $j(i)^* = j(i)$, where $j(i)^*$ denotes the length of $\boldsymbol{\delta}^*$. We obtain

$$\begin{aligned} F\left(\mathbf{0}_{j(i)}\right) &= \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta_{j(i)}), \\ F\left(\mathbf{0}_{j(i)^*}\right) &= \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta_{j(i)}). \end{aligned}$$

It holds that the first $j(i) - 1$ entries in the product are equal. For the $j(i)$-th (last) entry it holds that: $\vartheta_{j(i)^*} \geq \vartheta_j$. This inequality stems from:

(i) In the DT approach, both parameters $\vartheta_{j(i)^*}$ and $\vartheta_{j(i)}$ are equal since the hazard rate for the i-th individual is constant within the discrete time intervals. Therefore, $\vartheta_{j(i)^*} = \vartheta_{j(i)}$,

which leads to $f(0 \mid \vartheta_{j(i)^*}) = f(0 \mid \vartheta_{j(i)})$ and from this, it follows that: $F\left(\mathbf{0}_{j(i)^*}\right) = F\left(\mathbf{0}_{j(i)}\right)$.

(ii) In the *PW* approach, the distribution parameter of the Poisson likelihood is defined as $\vartheta_j = \exp(\eta_j)\exp(o_j)$, where o_j is the offset. Since $t < t^*$, it holds that $\vartheta^*_{j(i)} > \vartheta_{j(i)}$, where $\vartheta^*_{j(i)}$ corresponds to t^* and $\vartheta_{j(i)}$ corresponds to t. Since

$$\vartheta_{j(i)} = \exp(\eta_{j(i)})\exp(o_{j(i)}) < \exp(\eta_{j(i)})\exp(o^*_{j(i)}) = \vartheta^*_{j(i)},$$

with $o^*_{j(i)} = o_{j(i)} + c,\ c > 0$. In other words, the offset o^*_j corresponding to t^* is larger than that of t, since t^* occurred later in the same time interval, therefore, $f(0 \mid \vartheta_{j(i)}) > f(0 \mid \vartheta_{j(i)^*})$, which follows from:

$$f(0 \mid \vartheta_{j(i)}) = \frac{\vartheta^0_{j(i)}\exp(-\vartheta_{j(i)})}{0!} = \exp(-\vartheta_{j(i)}),\ \text{and}$$
$$f(0 \mid \vartheta^*_{j(i)}) = \frac{{\vartheta^*_{j(i)}}^0\exp(-\vartheta^*_{j(i)})}{0!} = \exp(-\vartheta^*_{j(i)}).$$

This holds because for any $a, b \in \mathbb{R}$ with $a < b \Rightarrow -a > -b \Rightarrow \exp(-a) > \exp(-b)$.

Case 2: $\delta = 0$ and $t \in (\kappa_{j-1},\ \kappa_j]$, $t^* \in (\kappa_{j+h},\ \kappa_{j+h+1}]$ (PW) or $t \in [\kappa_{j-1},\ \kappa_j)$, $t^* \in [\kappa_{j+h},\ \kappa_{j+h+1})$ (DT), with $h = 0, 1, 2, 3, \ldots$. In other words, t and t^* lie in different, possibly neighbouring, intervals. This means that $j(i)^* - j(i) \geq 1$. In this case we get:

$$F\left(\mathbf{0}_{j(i)}\right) = \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta_{j(i)}),$$
$$F\left(\mathbf{0}_{j(i)^*}\right) = \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta^*_{j(i)}) \prod_{j=j(i)+1}^{j(i)^*} f(0 \mid \vartheta_j),$$

with $\vartheta^*_{j(i)}$ from Case 1. The difference between $F\left(\mathbf{0}_{j(i)}\right)$ and $F\left(\mathbf{0}_{j(i)^*}\right)$ is then:

$$F\left(\mathbf{0}_{j(i)}\right) - F\left(\mathbf{0}_{j(i)^*}\right) = \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta_{j(i)}) - \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(0 \mid \vartheta^*_{j(i)}) \prod_{j=j(i)+1}^{j(i)^*} f(0 \mid \vartheta_j),$$

Since the first $j(i) - 1$ entries are the same, we write them as a common factor:

$$= \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) \left[f(0|\vartheta_{j(i)}) - \underbrace{f(0|\vartheta^*_{j(i)})}_{\substack{< f(0|\vartheta_{j(i)}),\\ \text{see Case 1}}} \underbrace{\prod_{j=j(i)+1}^{j(i)^*} f\left(0|\vartheta_j\right)}_{<1} \right].$$

Thus, the difference is at least zero: $F\left(\mathbf{0}_{j(i)}\right) - \tilde{F}\left(\mathbf{0}_{j(i)^*}\right) \geq 0$, proving Case 2.

Case 3: $\delta = 1$ and $t, t^* \in (\kappa_{j-1},\ \kappa_j]$ (PW) or $t, t^* \in [\kappa_{j-1},\ \kappa_j)$ (DT). In other words, t, t^* lie

in the same interval ($j(i)^* = j(i)$). In this case we get:

$$\begin{aligned} F\left((\mathbf{0}_{j(i)-1}, 1)\right) &= f\left(\mathbf{0}_{j(i)}\right) + f\left((\mathbf{0}_{j(i)-1}, 1)\right) \\ &= \prod_{j=1}^{j(i)} f(0 \mid \vartheta_j) + \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)}), \quad \text{and} \\ F\left((\mathbf{0}_{j(i)^*-1}, 1)\right) &= f\left(\mathbf{0}_{j(i)^*}\right) + f\left((\mathbf{0}_{j(i)^*-1}, 1)\right) \\ &= \prod_{j=1}^{j(i)} f(0 \mid \vartheta_j) + \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)^*}). \end{aligned}$$

The difference is in this case at least zero, i.e.

$$F\left(\left(\mathbf{0}_{j(i)-1}, 1\right)\right) - F\left(\left(\mathbf{0}_{j(i)^*-1}, 1\right)\right) \geq 0.$$

This is due to:

(i) Recall that in the *DT* approach, $f(0|\vartheta_{j(i)}) = f(0|\vartheta_{j(i)^*})$ and $f(1|\vartheta_{j(i)}) = f(1|\vartheta_{j(i)^*})$. Hence, the difference will be zero.

(ii) In the *PW* approach, we already established that $f(0|\vartheta_{j(i)}) > f(0|\vartheta_{j(i)^*})$ and recall that the first $j(i) - 1$ entries are the same. We can write the difference as:

$$\begin{aligned} &F\left((\mathbf{0}_{j(i)-1}, 1)\right) - F\left((\mathbf{0}_{j(i)^*-1}, 1)\right) = \\ &\prod_{j=1}^{j(i)} f(0 \mid \vartheta_j) + \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)}) - \left[\prod_{j=1}^{j(i)^*} f(0 \mid \vartheta_j) + \prod_{j=1}^{j(i)^*-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)^*})\right] \\ &= \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j)\Big[f(0 \mid \vartheta_{j(i)}) + f(1 \mid \vartheta_{j(i)})\Big] - \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j)\Big[f(0 \mid \vartheta_{j(i)^*}) + f(1 \mid \vartheta_{j(i)^*})\Big]. \end{aligned}$$

The difference depends entirely on the individual likelihoods:

$$= \prod_{j=1}^{j(i)-1} f(0|\vartheta_j)\left[f(0|\vartheta_{j(i)}) + f(1|\vartheta_{j(i)}) - f(0|\vartheta_{j(i)^*}) - f(1|\vartheta_{j(i)^*})\right].$$

Ignoring the common factor, we are left with:

$$\propto f(0 \mid \vartheta_{j(i)}) - f(0 \mid \vartheta_{j(i)^*}) + f(1 \mid \vartheta_{j(i)}) - f(1 \mid \vartheta_{j(i)^*}).$$

Recall that the parameters of the respective Poisson likelihoods have the following relationship: $\vartheta_{j(i)} = \exp(\eta_{j(i)} + o_{j(i)}) < \exp(\eta_{j(i)} + o_{j(i)} + c) = \vartheta_{j(i)^*}$ with $c > 0$. Hence, we may write $\vartheta_{j(i)} < (\vartheta_{j(i)} + \epsilon) = \vartheta_{j(i)^*}$ with $\epsilon > 0$. We plug-in the individual Poisson likelihoods with their

respective parameters:

$$\begin{aligned}
&\propto \underbrace{\exp\left(-\vartheta_{j(i)}\right)}_{=f(0|\vartheta_{j(i)})} - \underbrace{\exp\left(-(\vartheta_{j(i)}+\epsilon)\right)}_{=f(0|\vartheta_{j(i)^*})} + \underbrace{\exp\left(-\vartheta_{j(i)}\right)\vartheta_{j(i)}}_{=f(1|\vartheta_{j(i)})} - \underbrace{\exp\left(-(\vartheta_{j(i)}+\epsilon)\right)\left(\vartheta_{j(i)}+\epsilon\right)}_{=f(1|\vartheta_{j(i)^*})} \\
&= \exp\left(-\vartheta_{j(i)}\right)\left(1-\exp(-\epsilon)\right) + \vartheta_{j(i)}\exp\left(-\vartheta_{j(i)}\right)\left(1-\exp(-\epsilon)\right) - \epsilon\exp(-(\vartheta_{j(i)}+\epsilon)) \\
&= (1-\exp(-\epsilon))\exp(-\vartheta_{j(i)})(1+\vartheta_{j(i)}) - \epsilon\exp(-(\vartheta_{j(i)}+\epsilon)) =: \Xi(\epsilon).
\end{aligned}$$

We now consider $\Xi(\epsilon)$ as a function of ϵ, i.e. the difference in the respective parameters $\vartheta_{j(i)}$ and $\vartheta_{j(i)^*}$. Our objective is to show that $\Xi(\epsilon)$ is a non-negative function for all values of $\epsilon \geq 0$. A sufficient condition for a function f being non-negative is: The function $f(x) > 0$ and for all x in $[a, b]$, $\partial f(x)/\partial x$ exists and is equal to a nonzero number. Here we consider all x in $[0, \infty)$, since $\epsilon < 0$ is not possible by construction. The first derivative of $\Xi(\epsilon)$ w.r.t. ϵ is equal to:

$$\begin{aligned}
\frac{\partial \Xi(\epsilon)}{\partial \epsilon} &= \exp(-\epsilon)\exp(-\vartheta_{j(i)})\left(1+\vartheta_{j(i)}\right) - \exp(-(\vartheta_{j(i)}+\epsilon)) + \epsilon\exp(-(\vartheta_{j(i)}+\epsilon)) \\
&= \exp(-(\vartheta_{j(i)}+\epsilon)) + \vartheta_{j(i)}\exp(-(\vartheta_{j(i)}+\epsilon)) - \exp(-(\vartheta_{j(i)}+\epsilon)) + \epsilon\exp(-(\vartheta_{j(i)}+\epsilon)) \\
&= \underbrace{\vartheta_{j(i)}\exp(-(\vartheta_{j(i)}+\epsilon)) + \epsilon\exp(-(\vartheta_{j(i)}+\epsilon))}_{\geq 0}.
\end{aligned}$$

By respectively setting $\epsilon = 0$ and $\epsilon = b > 0$, it can be seen that:

$$\begin{aligned}
\left.\frac{\partial \Xi(\epsilon)}{\partial \epsilon}\right|_{\epsilon=0} &= \vartheta_{j(i)}\exp(-\vartheta_{j(i)}) > 0, \\
\left.\frac{\partial \Xi(\epsilon)}{\partial \epsilon}\right|_{\epsilon=b} &= \vartheta_{j(i)}\exp(-(\vartheta_{j(i)}+b)) + b\exp(-(\vartheta_{j(i)}+b)) > 0.
\end{aligned}$$

Since $\Xi(0) = 0$, i.e. the beginning of the interval $[0, \infty)$, and the first derivative is non-negative for any $x \in [0, \infty)$, we can conclude that $\Xi(\epsilon)$ is a non-negative function. Hence, the difference is at least zero for the PW approach as well.

Case 4: $\delta = 1$ and $t \in (\kappa_{j-1},\ \kappa_j]$, $t^* \in (\kappa_{j+h},\ \kappa_{j+h+1}]$ (PW) or $t \in [\kappa_{j-1},\ \kappa_j)$, $t^* \in [\kappa_{j+h},\ \kappa_{j+h+1})$ (DT), with $h = 0, 1, 2, 3, \ldots$. This implies that $\tilde{j}(i) - j(i) \geq 1$. We get:

$$\begin{aligned}
F\left((\mathbf{0}_{j(i)-1}, 1)\right) &= f\left(\mathbf{0}_{j(i)}\right) + f\left((\mathbf{0}_{j(i)-1}, 1)\right) \\
&= \prod_{j=1}^{j(i)} f(0|\vartheta_j) + \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)}) \\
&= \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) f(0|\vartheta_{j(i)}) + \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)}) \\
&= \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) \left[f\left(0|\vartheta_{j(i)}\right) + f\left(1|\vartheta_{j(i)}\right) \right],
\end{aligned}$$

$$
\begin{aligned}
F\left((\mathbf{0}_{j(i)^*-1},1)\right) &= f\left(\mathbf{0}_{j(i)^*}\right) + f\left((\mathbf{0}_{j(i)^*-1},1)\right) \\
&= \prod_{j=1}^{j(i)^*} f(0|\vartheta_j) + \prod_{j=1}^{j(i)^*-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)^*}) \\
&= \prod_{j=1}^{j(i)^*-1} f(0|\vartheta_j) f(0|\vartheta_{j(i)^*}) \prod_{j=j(i)+1}^{j(i)^*} f(0|\vartheta_j) + \prod_{j=1}^{j(i)-1} f(0|\vartheta_j) f(0|\vartheta_{j(i)^*}) \prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)^*}) \\
&= \prod_{j=1}^{j(i)^*-1} f(0|\vartheta_j) \left[f(0|\vartheta_{j(i)^*}) \prod_{j=j(i)+1}^{j(i)^*} f(0|\vartheta_j) + f(0|\vartheta_{j(i)^*}) \prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)^*}) \right] \\
&= \prod_{j=1}^{j(i)^*-1} f(0|\vartheta_j) \left[f(0|\vartheta_{j(i)^*}) \left(\prod_{j=j(i)+1}^{j(i)^*} f(0|\vartheta_j) + \prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)^*}) \right) \right].
\end{aligned}
$$

The difference is at least zero for DT and PW approaches:

$$
\begin{aligned}
&F\left((\mathbf{0}_{j(i)-1},1)\right) - F\left((\mathbf{0}_{j(i)^*-1},1)\right) = \\
&= \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) \left[f(0 \mid \vartheta_{j(i)}) + f(1 \mid \vartheta_{j(i)}) \right] - \\
&\prod_{j=1}^{j(i)^*-1} f(0|\vartheta_j) \left[f(0|\vartheta_{j(i)^*}) \left(\prod_{j=j(i)+1}^{j(i)^*} f(0|\vartheta_j) + \prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j) f(1|\vartheta_{j(i)^*}) \right) \right] \\
&= \prod_{j=1}^{j(i)-1} f(0 \mid \vartheta_j) \Bigg[f(\ 0 \mid \vartheta_{j(i)}) + f(1 \mid \vartheta_{j(i)}) - \\
&f(0 \mid \vartheta_{j(i)^*}) \left(\prod_{j=j(i)+1}^{j(i)^*} f(0 \mid \vartheta_j) + \prod_{j=j(i)+1}^{j(i)^*-1} f(0 \mid \vartheta_j) f(1 \mid \vartheta_{j(i)^*}) \right) \Bigg].
\end{aligned}
$$

Ignoring the common factor that contains all $j(i)-1$ terms leads to:

$$
\propto f(0 \mid \vartheta_{j(i)}) + f(1 \mid \vartheta_{j(i)}) - f(0 \mid \vartheta_{j(i)^*}) \underbrace{\left(\prod_{j=j(i)+1}^{j(i)^*} f(0|\vartheta_j) + \underbrace{\prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j)}_{=:\Upsilon} f(1|\vartheta_{j(i)^*}) \right)}_{=:(**)<1}.
$$

Using once again $f(0|\vartheta_{j(i)}) > f(0|\vartheta_{j(i)^*})$ completes the proof:

$$
\begin{aligned}
&\underbrace{f(0|\vartheta_{j(i)})}_{>0} - \underbrace{f(0|\vartheta_{j(i)^*})}_{<f(0|\vartheta_{j(i)})} \underbrace{\left[\prod_{j=j(i)+1}^{j(i)^*} f(\delta_j = 0|\vartheta_j) + \underbrace{\prod_{j=j(i)+1}^{j(i)^*-1} f(0|\vartheta_j)}_{=:\Upsilon} f(1|\vartheta_{j(i)^*}) \right]}_{=:(**)<1} + \underbrace{f(1|\vartheta_{j(i)})}_{>0} \qquad \text{(SA1)} \\
&= \underbrace{f(0|\vartheta_{j(i)}) - f(0|\vartheta_{j(i)^*})(**)}_{>0} + \underbrace{f(1|\vartheta_{j(i)})}_{\geq 0} > 0.
\end{aligned}
$$

Where $(**)$ follows from the fact that the individual probabilities are from incomplete supports of Poisson densities that are required for the likelihoods. Therefore, the individual evaluations will not add up to 1. Additionally, the quantities in $(**)$ are multiplied by a term that is

smaller than $f(0|\vartheta_{j(i)})$. Note that in the case that t and t^* lie in subsequent intervals (i.e. next to each other), the term Υ in Equation (SA1) would be a product over an empty set, since t and t^* lie in intervals next to each other. In this case Υ would equal to a factor of zero. Nevertheless, this does not affect the proof, since the sum in $(**)$ would involve evaluations of probability mass functions of a Poisson distribution over two points in its domain ($\delta_{j(i)^*} = 0$ and $\delta_{j(i)^*} = 1$). Since the Poisson distribution has infinite support on the positive integers, the term $(**)$ will result in a value below 1. Thus, we have proven that the proposed function $F(\boldsymbol{\delta})$ is a monotonically decreasing function of the original event-time random variable for both DT and PW approaches. Hence, replacing the survival function of T denoted by $S(t)$ with the proposed function $F(\boldsymbol{\delta})$ in a bivariate copula model preserves the original dependence structure of interest.

Part B: Details of DT and PW functions

B1 General definition of $F(\boldsymbol{\delta})$ and $f(\boldsymbol{\delta})$ via DT approach

Consider a suitable link function $g(\cdot)$ such as logit, probit, clog-log, and log-log with corresponding response function $g^{-1}(\cdot)$, i.e. $\vartheta_j = g^{-1}(\eta_j) \in [0,1]$, with $\eta_j = \mathbf{x}_j^\top \boldsymbol{\beta}$. The *DT* function based on the Bernoulli likelihood is then:

$$f(\boldsymbol{\delta}) = \begin{cases} \prod_{j=1}^{j(i)} \left(1 - g^{-1}(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}, \\ \prod_{j=1}^{j(i)-1} \left(1 - g^{-1}(\eta_j)\right) g^{-1}(\eta_{j(i)}), & \text{if } \boldsymbol{\delta} = (\mathbf{0}_{j(i)-1},\ 1). \end{cases} \tag{B1}$$

By applying the generalised product rule for differentiation to Equation (B1), one obtains the first order partial derivatives w.r.t. $\boldsymbol{\beta}$. This is equivalent to writing:

$$\begin{aligned} f(\mathbf{0}_{j(i)}) &= \prod_{j=1}^{j(i)} \left(1 - g^{-1}(\eta_j)\right) = \exp\left(\sum_{j=1}^{j(i)} \log\left(1 - g^{-1}(\eta_j)\right)\right), \\ f((\mathbf{0}_{j(i)-1},\ 1)) &= \prod_{j=1}^{j(i)-1} \left(1 - g^{-1}(\eta_j)\right) g^{-1}(\eta_{j(i)}) \\ &= \exp\left(\sum_{j=1}^{j(i)-1} \log\left(1 - g^{-1}(\eta_j)\right) + \log\left(g^{-1}(\eta_{j(i)})\right)\right), \end{aligned}$$

and then applying the chain rule w.r.t. $\boldsymbol{\beta}$ in order to obtain:

$$\frac{\partial f(\mathbf{0}_{j(i)})}{\partial \boldsymbol{\beta}} = \left(\prod_{j=1}^{j(i)} \left(1 - g^{-1}(\eta_j)\right)\right) \left[\sum_{j=1}^{j(i)} \frac{-\frac{\partial g^{-1}(\eta_j)}{\partial \eta_j}}{1 - g^{-1}(\eta_j)} \mathbf{x}_j\right],$$

as well as

$$\frac{\partial f((\mathbf{0}_{j(i)-1},\ 1))}{\partial \boldsymbol{\beta}} = \left(\prod_{j=1}^{j(i)-1} \left(1 - g^{-1}(\eta_j)\right) g^{-1}(\eta_{j(i)})\right) \left[\sum_{j=1}^{j(i)-1} \left(\frac{-\frac{\partial g^{-1}(\eta_j)}{\partial \eta_j}}{1 - g^{-1}(\eta_j)} \mathbf{x}_j\right) + \frac{\frac{\partial g^{-1}(\eta_{j(i)})}{\partial \eta_{j(i)}}}{g^{-1}(\eta_{j(i)})} \mathbf{x}_{j(i)}\right].$$

The second order partial derivatives are:

$$\begin{aligned} \frac{\partial^2 f(\mathbf{0}_{j(i)})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^\top} = \left(\prod_{j=1}^{j(i)} \left(1 - g^{-1}(\eta_j)\right)\right) \Bigg\{ &\left[\sum_{j=1}^{j(i)} \frac{-\frac{\partial g^{-1}(\eta_j)}{\partial \eta_j}}{1 - g^{-1}(\eta_j)} \mathbf{x}_j\right] \left[\sum_{j=1}^{j(i)} \frac{-\frac{\partial g^{-1}(\eta_j)}{\partial \eta_j}}{1 - g^{-1}(\eta_j)} \mathbf{x}_j\right]^\top + \\ &\sum_{j=1}^{j(i)} \left(\frac{-\frac{\partial^2 g^{-1}(\eta_j)}{\partial \eta_j^2} \left(1 - g^{-1}(\eta_j)\right) \mathbf{x}_j \mathbf{x}_j^\top - \left(\frac{\partial g^{-1}(\eta_j)}{\partial \eta_j}\right)^2 \mathbf{x}_j \mathbf{x}_j^\top}{\left(1 - g^{-1}(\eta_j)\right)^2}\right)\Bigg\}, \end{aligned}$$

and

$$\frac{\partial^2 f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \left(\prod_{j=1}^{j(i)-1}\left(1-g^{-1}\left(\eta_j\right)\right)g^{-1}\left(\eta_j\right)\right)$$
$$\left\{\left[\sum_{j=1}^{j(i)-1}\frac{-\frac{\partial g^{-1}(\eta_j)}{\partial\eta_j}}{1-g^{-1}\left(\eta_j\right)}\mathbf{x}_j+\frac{\frac{\partial g^{-1}\left(\eta_{j(i)}\right)}{\partial\eta_{j(i)}}}{g^{-1}\left(\eta_{j(i)}\right)}\mathbf{x}_{j(i)}\right]\left[\sum_{j=1}^{j(i)-1}\frac{-\frac{\partial g^{-1}(\eta_j)}{\partial\eta_j}}{1-g^{-1}\left(\eta_j\right)}\mathbf{x}_j+\frac{\frac{\partial g^{-1}\left(\eta_{j(i)}\right)}{\partial\eta_{j(i)}}}{g^{-1}\left(\eta_{j(i)}\right)}\mathbf{x}_{j(i)}\right]^\top+$$

$$\left[\sum_{j=1}^{j(i)-1}\left(1-g^{-1}\left(\eta_j\right)\right)^{-2}\left(-\frac{\partial^2 g^{-1}\left(\eta_j\right)}{\partial\eta_j^2}\left(1-g^{-1}\left(\eta_j\right)\right)\mathbf{x}_j\mathbf{x}_j^\top-\left(\frac{\partial g^{-1}\left(\eta_j\right)}{\partial\eta_j}\right)^2\mathbf{x}_j\mathbf{x}_j^\top\right)+\right.$$
$$\left.\left(g^{-1}\left(\eta_{j(i)}\right)\right)^{-2}\left(\frac{\partial^2 g^{-1}\left(\eta_{j(i)}\right)}{\partial\eta_{j(i)}^2}\left(g^{-1}\left(\eta_{j(i)}\right)\right)\mathbf{x}_{j(i)}\mathbf{x}_{j(i)}^\top-\left(\frac{\partial g^{-1}\left(\eta_{j(i)}\right)}{\partial\eta_{j(i)}}\right)^2\mathbf{x}_{j(i)}\mathbf{x}_{j(i)}^\top\right)\right]\Bigg\}$$

The individual first and second order partial derivatives of the response functions $g^{-1}(\cdot)$ w.r.t. the predictor η_j are known from generalised linear model methodology.

B2 Definition of $F(\boldsymbol{\delta})$ and $f(\boldsymbol{\delta})$ via DT approach with clog-log link function

The first and second order partial derivatives of the proposed functions $F(\boldsymbol{\delta})$ and $f(\boldsymbol{\delta})$ constructed using the *DT* approach w.r.t. a coefficient vector $\boldsymbol{\beta}$ are derived as follows: Recall the *DT* setting with generic structured additive predictor and response function set to the inverse of the clog-log link, i.e. $\vartheta_j = g^{-1}(\eta_j) = 1-\exp\left(-\exp\left(\eta_j\right)\right)$ and $\eta_j = \mathbf{x}_j^\top\boldsymbol{\beta}$. The function using the *DT* approach is given by:

$$f\left(\boldsymbol{\delta}\right) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)}\exp\left(\eta_j\right)\right), & \text{if } \boldsymbol{\delta}=\mathbf{0}_{j(i)}, \\ \exp\left(-\sum_{j=1}^{j(i)-1}\exp\left(\eta_j\right)\right)-\exp\left(-\sum_{j=1}^{j(i)}\exp\left(\eta_j\right)\right), & \text{if } \boldsymbol{\delta}=\left(\mathbf{0}_{j(i)-1},\ 1\right). \end{cases} \tag{B2}$$

By applying the chain rule to Equation (B2) one obtains:

$$\frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}} = -\exp\left(-\sum_{j=1}^{j(i)}\exp\left(\eta_j\right)\right)\sum_{j=1}^{j(i)}\left[\exp\left(\eta_j\right)\mathbf{x}_j\right],$$
$$\frac{\partial f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}} = -\exp\left(-\sum_{j=1}^{j(i)-1}\exp\left(\eta_j\right)\right)\left[\sum_{j=1}^{j(i)-1}\exp\left(\eta_j\right)\mathbf{x}_j\right]+\exp\left(-\sum_{j=1}^{j(i)}\exp\left(\eta_j\right)\right)\left[\sum_{j=1}^{j(i)}\exp\left(\eta_j\right)\mathbf{x}_j\right].$$

The first order partial derivatives of $F(\boldsymbol{\delta})$ are equal to:

$$\frac{\partial F\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}} = \frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}}, \qquad \frac{\partial F\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}} = \frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}}+\frac{\partial f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}}.$$

The second order partial derivatives of $f(\boldsymbol{\delta})$ w.r.t. $\boldsymbol{\beta}$ are given by:

$$\frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right)\left\{\left[\sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\right]\left[\sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\right]^\top - \sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\mathbf{x}_j^\top\right\},$$

$$\frac{\partial^2 f\left(\left(\mathbf{0}_{j(i)-1},1\right)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \exp\left(-\sum_{j=1}^{j(i)-1}\exp(\eta_j)\right)\left\{\left[\sum_{j=1}^{j(i)-1}\exp(\eta_j)\mathbf{x}_j\right]\left[\sum_{j=1}^{j(i)-1}\exp(\eta_j)\mathbf{x}_j\right]^\top - \sum_{j=1}^{j(i)-1}\exp(\eta_j)\mathbf{x}_j\mathbf{x}_j^\top\right\} -$$
$$\exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right)\left\{\left[\sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\right]\left[\sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\right]^\top - \sum_{j=1}^{j(i)}\exp(\eta_j)\mathbf{x}_j\mathbf{x}_j^\top\right\}.$$

Finally, the second order partial derivatives of $F(\cdot)$ w.r.t. $\boldsymbol{\beta}$ are obtained in the same fashion as the first order partial derivatives, i.e.

$$\frac{\partial^2 F\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}, \qquad \frac{\partial^2 F\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} + \frac{\partial^2 f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}.$$

B3 Definition of $F(\boldsymbol{\delta})$ and $f(\boldsymbol{\delta})$ via PW approach

The first and second order partial derivatives of the proposed functions $F(\boldsymbol{\delta})$ and $f(\boldsymbol{\delta})$ constructed using the PW approach w.r.t. a coefficient vector $\boldsymbol{\beta}$ are derived as follows: Recall the PW function defined in Equation (5) with generic structured additive predictor and exponential response function, i.e. $\vartheta_j = g^{-1}(\eta_j) = \exp(\eta_j)$ and $\eta_j = \mathbf{x}_j^\top\boldsymbol{\beta}$:

$$f(\boldsymbol{\delta}) = \begin{cases} \exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right), & \text{if } \boldsymbol{\delta} = \mathbf{0}_{j(i)}, \\ \exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right)\exp\left(\eta_{j(i)}\right), & \text{if } \boldsymbol{\delta} = \left(\mathbf{0}_{j(i)-1},\ 1\right). \end{cases} \tag{B3}$$

By applying the chain rule to Equation (B3) one obtains:

$$\frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}} = -\exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right)\sum_{j=1}^{j(i)}\left[\exp(\eta_j)\,\mathbf{x}_j\right],$$
$$\frac{\partial f\left(\left(\mathbf{0}_{j(i)-1},1\right)\right)}{\partial\boldsymbol{\beta}} = \exp\left(-\sum_{j=1}^{j(i)}\left[\exp(\eta_j)\right] + \eta_{j(i)}\right)\left[-\sum_{j=1}^{j(i)}\left[\exp(\eta_j)\,\mathbf{x}_j\right] + \mathbf{x}_{j(i)}\right].$$

Then, the first order partial derivatives of $F(\boldsymbol{\delta})$ are:

$$\frac{\partial F\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}} = \frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}}, \qquad \frac{\partial F\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}} = \frac{\partial f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}} + \frac{\partial f\left(\left(\mathbf{0}_{j(i)-1},\ 1\right)\right)}{\partial\boldsymbol{\beta}}.$$

The second order partial derivatives of $f(\boldsymbol{\delta})$ w.r.t. $\boldsymbol{\beta}$ are given by:

$$\frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \exp\left(-\sum_{j=1}^{j(i)}\exp(\eta_j)\right)\left\{\sum_{j=1}^{j(i)}[\exp(\eta_j)\mathbf{x}_j]\sum_{j=1}^{j(i)}[\exp(\eta_j)\mathbf{x}_j]^\top - \sum_{j=1}^{j(i)}\left[\exp(\eta_j)\mathbf{x}_j\mathbf{x}_j^\top\right]\right\},$$

$$\frac{\partial^2 f\left((\mathbf{0}_{j(i)-1},1)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \exp\left(-\sum_{j=1}^{j(i)}[\exp(\eta_j)] + \eta_{j(i)}\right)\left\{\left[-\sum_{j=1}^{j(i)}[\exp(\eta_j)\mathbf{x}_j] + \mathbf{x}_{j(i)}\right]\left[-\sum_{j=1}^{j(i)}[\exp(\eta_j)\mathbf{x}_j] + \mathbf{x}_{j(i)}\right]^\top - \sum_{j=1}^{j(i)}\left[\exp(\eta_j)\mathbf{x}_j\mathbf{x}_j^\top\right]\right\}.$$

Finally, the second order partial derivatives of $F(\cdot)$ w.r.t. $\boldsymbol{\beta}$ are obtained in the same fashion as the first order partial derivatives, i.e.

$$\frac{\partial^2 F\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}, \qquad \frac{\partial^2 F\left((\mathbf{0}_{j(i)-1},\ 1)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} = \frac{\partial^2 f\left(\mathbf{0}_{j(i)}\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top} + \frac{\partial^2 f\left((\mathbf{0}_{j(i)-1},\ 1)\right)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^\top}.$$

B4 Partial derivatives of the model's log-likelihood

The log-likelihood of the i-th observation with $i = 1, \ldots n$ is given by:

$$\ell_i = (1-y_i)\ln\Big\{C[F_{1i}, F_{2i};\ \vartheta_i^{(c)}] - C[F_{1i}, F_{2i} - f_{2i};\ \vartheta_i^{(c)}]\Big\} + y_i \ln\Big\{f_{2i} - C[F_{1i}, F_{2i};\ \vartheta_i^{(c)}] + C[F_{1i}, F_{2i} - f_{2i};\ \vartheta_i^{(c)}]\Big\},$$

where $F_{1i} = F_1(0|\vartheta_i^{(1)})$, $F_{2i} = F_2(\boldsymbol{\delta}_i|\boldsymbol{\vartheta}_i^{(2)})$ and $f_{2i} = f_2(\boldsymbol{\delta}_i|\boldsymbol{\vartheta}_i^{(2)})$. The log-likelihood of the sample is obtained by summing up the individual log-likelihood contributions $\ell = \sum_{i=1}^n \ell_i$. We use the abbreviation: $\Delta_i = C[F_{1i}, F_{2i};\ \vartheta_i^{(c)}] - C[F_{1i}, F_{2i} - f_{2i};\ \vartheta_i^{(c)}]$. The first order partial derivatives are then:

$$\frac{\partial\ell_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(1)}} = \left[\frac{1-y_i}{\Delta_i} - \frac{y_i}{f_{2i}-\Delta_i}\right]\frac{\partial\Delta_i}{\partial F_{1i}}\frac{\partial F_{1i}}{\partial\boldsymbol{\beta}^{(1)}},$$

$$\frac{\partial\ell_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(2)}} = \left[\frac{1-y_i}{\Delta_i}\right]\left[\frac{\partial C(F_{1i}, F_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}}\frac{\partial F_{2i}}{\partial\boldsymbol{\beta}^{(2)}} - \frac{\partial C(F_{1i}, F_{2i} - f_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i} - f_{2i}}\left(\frac{\partial F_{2i}}{\partial\boldsymbol{\beta}^{(2)}} - \frac{\partial f_{2i}}{\partial\boldsymbol{\beta}^{(2)}}\right)\right] + \left[\frac{y_i}{f_{2i}-\Delta_i}\right]\left[\frac{\partial f_{2i}}{\partial\boldsymbol{\beta}^{(2)}} - \frac{\partial C(F_{1i}, F_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}}\frac{\partial F_{2i}}{\partial\boldsymbol{\beta}^{(2)}} + \frac{\partial C(F_{1i}, F_{2i} - f_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i} - f_{2i}}\left(\frac{\partial F_{2i}}{\partial\boldsymbol{\beta}^{(2)}} - \frac{\partial f_{2i}}{\partial\boldsymbol{\beta}^{(2)}}\right)\right],$$

$$\frac{\partial\ell_i(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}^{(c)}} = \left[\frac{1-y_i}{\Delta_i} - \frac{y_i}{f_{2i}-\Delta_i}\right]\frac{\partial\Delta_i}{\partial\vartheta_i^{(c)}}\frac{\partial\vartheta_i^{(c)}}{\partial\boldsymbol{\beta}^{(c)}}.$$

The first order partial derivatives with respect to the unknown regression coefficient vectors

$(\boldsymbol{\beta}^{(1)})^\top, (\boldsymbol{\beta}^{(2)})^\top, (\boldsymbol{\beta}^{(c)})^\top$ is then:

$$\boldsymbol{s}_i(\boldsymbol{\beta}) = \left(\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)}},\ \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)}},\ \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(c)}} \right).$$

The second order partial derivatives required for the Hessian $\boldsymbol{H}_i(\boldsymbol{\beta})$ are:

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(1)\top}} = \left[-\frac{1-y_{1i}}{\Delta_i^2} - \frac{y_{1i}}{(f_{2i}-\Delta_i)^2} \right] \left[\frac{\partial \Delta_i}{\partial F_{1i}} \frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \right] \left[\frac{\partial \Delta_i}{\partial F_{1i}} \frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \right]^\top + \left[\frac{1-y_{1i}}{\Delta_i} - \frac{y_{1i}}{f_{2i}-\Delta_i} \right] \left[\frac{\partial^2 \Delta_i}{\partial F_{1i}^2} \frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \left(\frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \right)^\top + \frac{\partial \Delta_i}{\partial F_{1i}} \frac{\partial^2 F_{1i}}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(1)\top}} \right]$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} = \left[-\frac{1-y_i}{\Delta_i^2} \right] \left[\frac{\partial \Delta_i}{\partial \boldsymbol{\beta}^{(2)}} \left(\frac{\partial \Delta_i}{\partial \boldsymbol{\beta}^{(2)}} \right)^\top \right] + \left[\frac{1-y_i}{\Delta_i} \right] \frac{\partial^2 \Delta_i}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} - \left[\frac{y_i}{(f_{2i}-\Delta_i)^2} \right] \left[\frac{\partial f_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial \Delta_i}{\partial \boldsymbol{\beta}^{(2)}} \right] \left[\frac{\partial f_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial \Delta_i}{\partial \boldsymbol{\beta}^{(2)}} \right]^\top + \left[\frac{y_i}{f_{2i}-\Delta_i} \right] \left[\frac{\partial^2 f_{2i}}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} - \frac{\partial^2 \Delta_i}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} \right],$$

with

$$\frac{\partial \Delta_i}{\partial \boldsymbol{\beta}^{(2)}} = \frac{\partial C(F_{1i}, F_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}} \frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial C(F_{1i}, F_{2i}-f_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}-f_{2i}} \left(\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial f_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right),$$

and

$$\frac{\partial^2 \Delta_i}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} = \frac{\partial^2 C(F_{1i}, F_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}^2} \frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \left(\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right)^\top + \frac{\partial C(F_{1i}, F_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}} \frac{\partial^2 F_{2i}}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} - \frac{\partial^2 C(F_{1i}, F_{2i}-f_{2i};\ \vartheta_i^{(c)})}{\partial (F_{2i}-f_{2i})^2} \left(\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial f_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right) \left(\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} - \frac{\partial f_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right)^\top + \frac{\partial C(F_{1i}, F_{2i}-f_{2i};\ \vartheta_i^{(c)})}{\partial F_{2i}-f_{2i}} \left(\frac{\partial^2 F_{2i}}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} - \frac{\partial^2 f_{2i}}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} \right).$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(c)} \partial \boldsymbol{\beta}^{(c)\top}} = \left[-\frac{1-y_i}{\Delta_i^2} - \frac{y_i}{(f_{2i}-\Delta_i)^2} \right] \left[\frac{\partial \Delta_i}{\partial \vartheta_i^{(c)}} \frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}} \right] \left[\frac{\partial \Delta_i}{\partial \vartheta_i^{(c)}} \frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}} \right]^\top + \left[\frac{1-y_i}{\Delta_i} - \frac{y_i}{f_{2i}-\Delta_i} \right] \left[\frac{\partial^2 \Delta_i}{\partial {\vartheta_i^{(c)}}^2} \frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}} \left(\frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}} \right)^\top + \frac{\partial \Delta_i}{\partial \vartheta_i^{(c)}} \frac{\partial^2 \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)} \partial \boldsymbol{\beta}^{(c)\top}} \right],$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(2)\top}} = \left[-\frac{1-y_i}{\Delta_i^2} + \frac{y_i}{(f_{2i}-\Delta_i)^2} \right] \frac{\partial \Delta_i}{\partial F_{1i}} \frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \left(\frac{\partial \Delta_i}{\partial F_{2i}} \frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right)^\top + \left[\frac{1-y_i}{\Delta_i} - \frac{y_i}{f_{2i}-\Delta_i} \right] \frac{\partial^2 \Delta_i}{\partial F_{1i}\, \partial F_{2i}} \frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}} \left(\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}} \right)^\top$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(c)\top}} = \left[-\frac{1-y_{1i}}{\Delta_i^2} - \frac{y_{1i}}{(f_{2i}-\Delta_i)^2}\right]\left[\frac{\partial \Delta_i}{\partial F_{1i}}\frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}}\right]\left[\frac{\partial \Delta_i}{\partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}}\right]^\top + \left[\frac{1-y_{1i}}{\Delta_i} - \frac{y_{1i}}{f_{2i}-\Delta_i}\right]\frac{\partial^2 \Delta_i}{\partial F_{1i}\,\partial \vartheta_i^{(c)}}\frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}}\left(\frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}}\right)^\top$$

$$\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(c)\top}} = \left[-\frac{1-y_{1i}}{\Delta_i^2} - \frac{y_{1i}}{(f_{2i}-\Delta_i)^2}\right]\left[\frac{\partial \Delta_i}{\partial F_{1i}}\frac{\partial F_{1i}}{\partial \boldsymbol{\beta}^{(1)}}\right]\left[\frac{\partial \Delta_i}{\partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}}\right]^\top + \left[\frac{1-y_{1i}}{\Delta_i} - \frac{y_{1i}}{f_{2i}-\Delta_i}\right]\frac{\partial^2 \Delta_i}{\partial F_{1i}\,\partial \vartheta_i^{(c)}}\frac{\partial F_{2i}}{\partial \boldsymbol{\beta}^{(2)}}\left(\frac{\partial \vartheta_i^{(c)}}{\partial \boldsymbol{\beta}^{(c)}}\right)^\top$$

These second order partial derivatives with respect to $\boldsymbol{\beta}^{(\bullet)}$, $\bullet \in \{1, 2, c\}$ are the blocks that constitute the Hessian:

$$\boldsymbol{H}_i(\boldsymbol{\beta}) = \begin{pmatrix} \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(1)\top}} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(2)\top}} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(c)\top}} \\ \left(\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(2)\top}}\right)^\top & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(2)\top}} & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(c)\top}} \\ \left(\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(1)} \partial \boldsymbol{\beta}^{(c)\top}}\right)^\top & \left(\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(2)} \partial \boldsymbol{\beta}^{(c)\top}}\right)^\top & \frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^{(c)} \partial \boldsymbol{\beta}^{(c)\top}} \end{pmatrix}$$

Lastly, $\boldsymbol{s}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{s}_i(\boldsymbol{\beta})$ and $\boldsymbol{H}(\boldsymbol{\beta}) = \sum_{i=1}^n \boldsymbol{H}_i(\boldsymbol{\beta})$

Part C: Additional empirical results

C1 Results assuming independent margins

Table C1: Estimated coefficients with 95% confidence intervals of the models assuming independent margins and using DT and PW approaches.

Binary margin `VisualSearch`				
$\beta_0^{(1)}$	-0.704	$[-0.849;\ -0.555]$		
`GenderMale`	0.174	[0.030; 0.318]		
`RiderTypeTakeAway`	0.501	[0.379; 0.627]		
`PositionMiddle`	0.276	[0.128; 0.424]		
`PositionCloseMotorLane`	0.338	[0.191; 0.488]		
		DT		PW
Time-to-event margin `WaitingTime`				
`GenderMale`	0.314	[0.148; 0.478]	0.291	[0.127; 0.456]
`RiderTypeTakeAway`	1.416	[1.248; 1.585]	1.459	[1.296; 1.627]
`PositionMiddle`	0.493	[0.326; 0.661]	0.481	[0.311; 0.656]
`PositionCloseMotorLane`	0.745	[0.584; 0.908]	0.729	[0.567; 0.897]
`GroupSize1to4`	-0.332	$[-0.456;\ -0.209]$	-0.361	$[-0.483;\ -0.240]$
`GroupSize5AndMore`	-0.790	$[-1.034;\ -0.548]$	-0.844	$[-1.087;\ -0.597]$
`ConformBehaviourYes`	0.267	[0.145; 0.387]	0.243	[0.118; 0.364]
Sample size: $n = 2{,}173$. Censoring rate: 47.78%.				

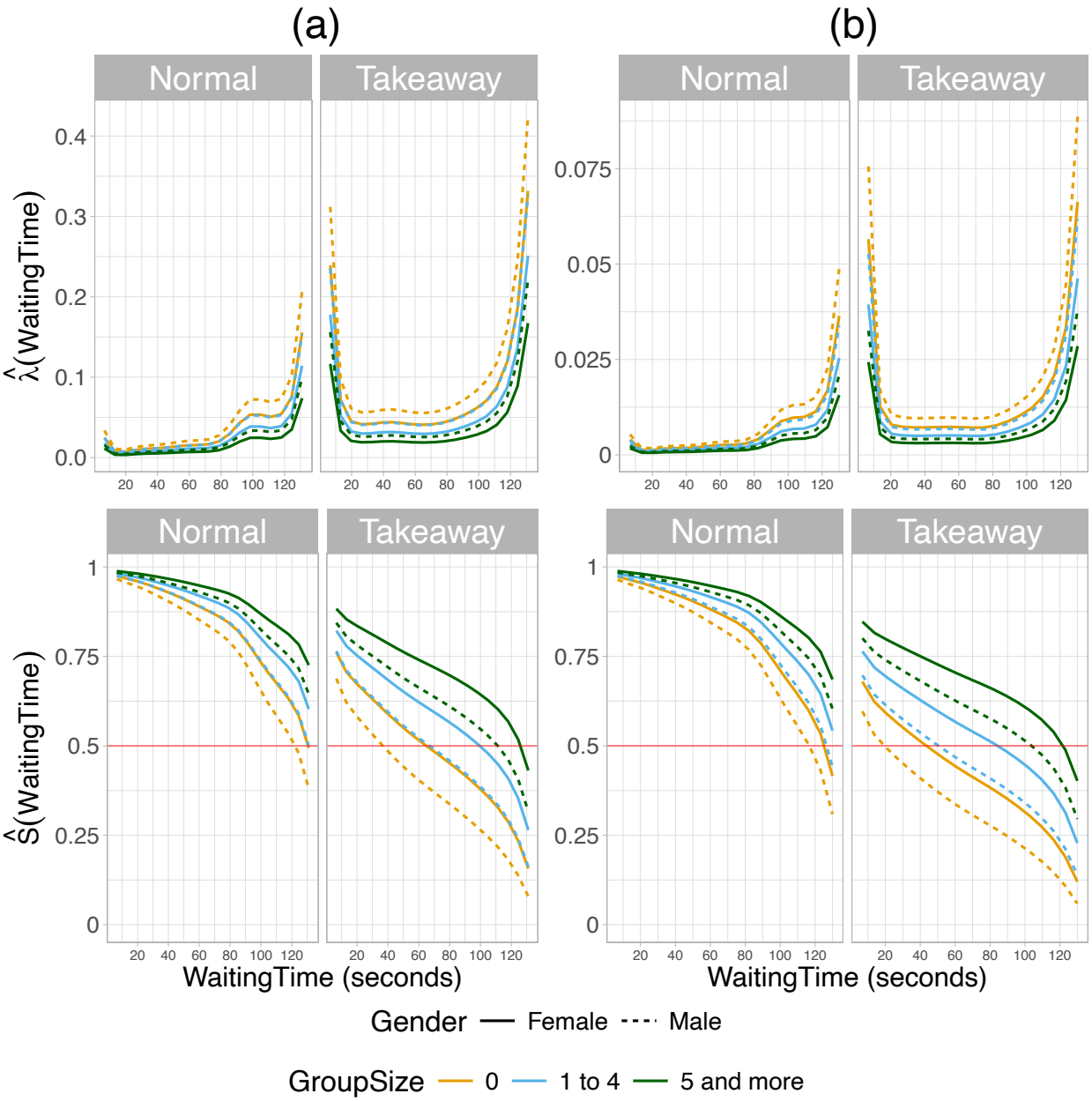

Figure C1: Estimated hazard rate and survival function across `RiderType`, `GroupSize` and `Gender` using independent univariate models with *DT* (a) and *PW* (b) approaches. Red line indicates the median.

C2 Additional diagnostics

We conduct an additional diagnostic check by splitting the data into the following four combinations:

$$\text{Combination 1: } (\texttt{RedLight} = 0;\ \texttt{VisualSearch} = 0),$$
$$\text{Combination 2: } (\texttt{RedLight} = 0;\ \texttt{VisualSearch} = 1),$$
$$\text{Combination 3: } (\texttt{RedLight} = 1;\ \texttt{VisualSearch} = 0),$$
$$\text{Combination 4: } (\texttt{RedLight} = 1;\ \texttt{VisualSearch} = 1).$$

Note that the four combinations can be made for the *DT* and *PW* approaches, respectively. After sorting the observations in the data according to these combinations, we transformed the margins evaluated at the CDF and the proposed function $F(\boldsymbol{\delta})$ to standard normal observations and plotted their contours using the function `VineCopula::BiCopKDE()` from the `VineCopula` package. The idea is to check whether the estimated contour plots resemble the shape of a Clayton copula. The estimated contours are shown in Figures C2 and C3. It can be seen that the contours from three out of the four combinations have a shape that look similar to a Clayton copula, albeit with some asymmetries to the lower-right corner. In case of the fourth combination (`RedLight` = 1; `VisualSearch` = 1), the contour shows a more complex shape that is unlikely to be approximated well by prominent parametric copula families, e.g. Clayton, or Frank. This complex shape obtained by observations in Combination 4 is more pronounced in Figure C3.

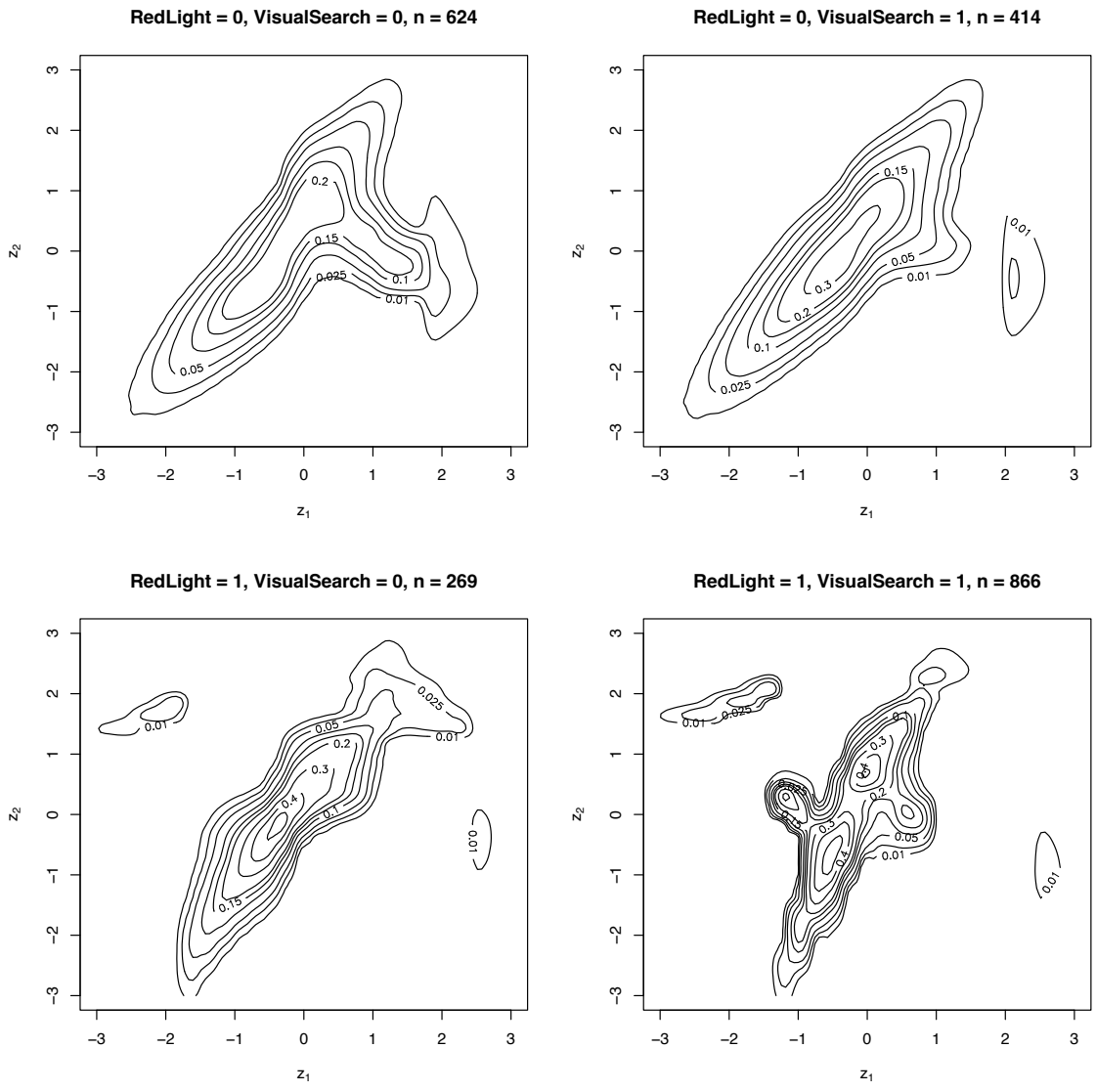

Figure C2: Estimated contours of the fitted margins transformed to standard normal margins using the *DT* approach, n indicates the number of observations in this combination.

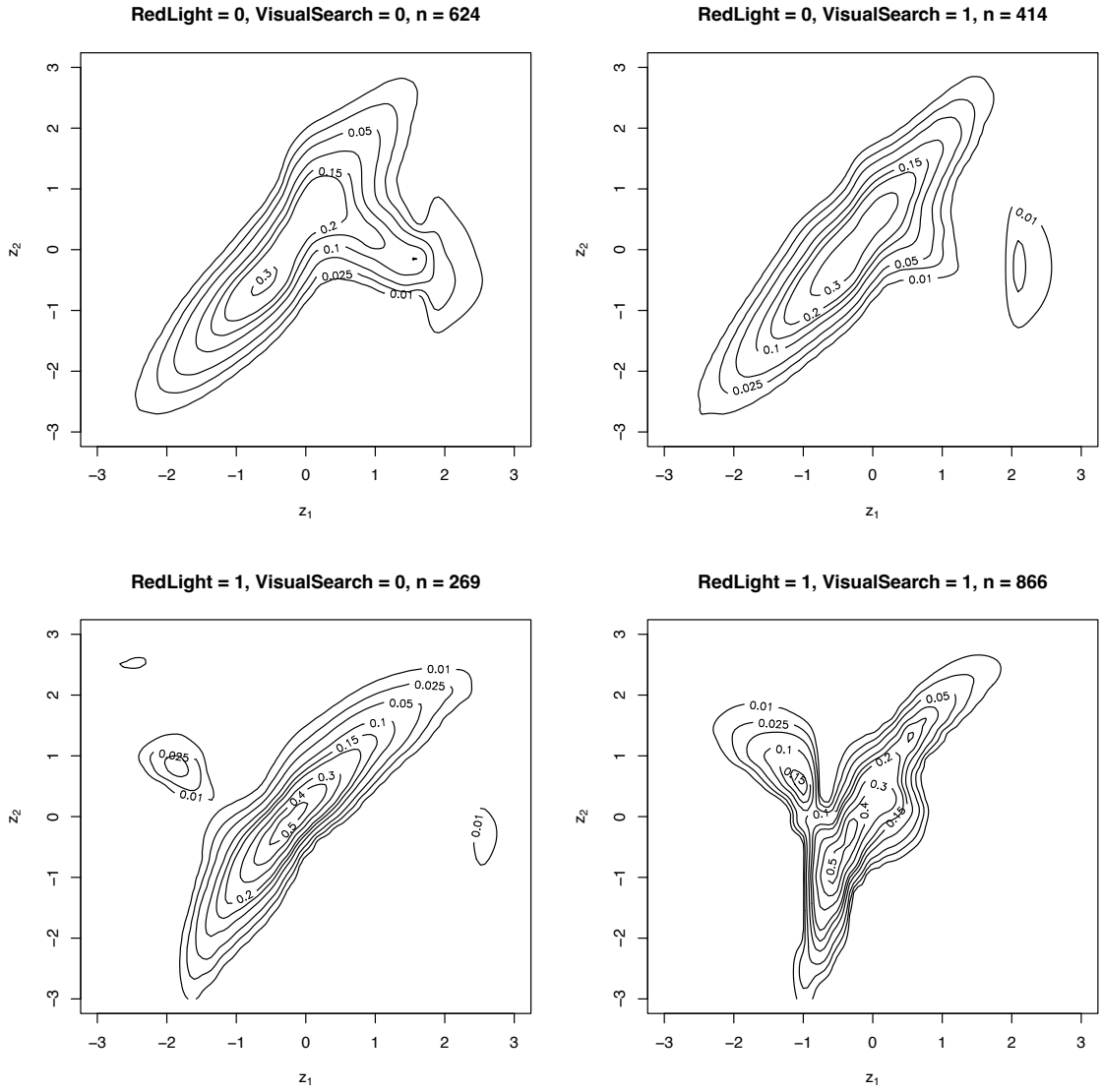

Figure C3: Estimated contours of the fitted margins transformed to standard normal margins using the *PW* approach, n indicates the number of observations in this combination.

Part D: Simulation study

We conduct various simulation studies in order to investigate the following questions:

Q1: Can the underlying baseline hazard rate, baseline cumulative hazard and baseline survival functions of the time-to-event margin be recovered?

Q2: Can the underlying effects of covariates on the respective parameters of the margins as well as the copula dependence parameter be recovered using the proposed distributional copula regression approach?

The simulation studies are based on structured additive predictors for all distribution parameters of the non-time-to-event margin, the survival function of the time-to-event variable, as well as the dependence parameter of the copula. We omit the index k in the distribution parameters for the remainder of this supplement due to both margins depending only on one parameter. The following structured additive predictors are used to generate the bivariate mixed response:

$$\begin{aligned}
\eta^{(1)} &= \beta_0^{(1)} \; + \; \beta_1^{(1)} x_1 \; + \; \beta_2^{(1)} x_2, && \text{(non-time-to-event margin)},\\
\eta^{(2)} &= s_0^{(2)}(t) \; + \; \beta_1^{(2)} x_2 \; + \; \beta_2^{(2)} x_4, && \text{(time-to-event margin)},\\
\eta^{(c)} &= \beta_0^{(c)} \; + \; \beta_1^{(c)} x_5, && \text{(copula parameter)},
\end{aligned}$$

where $s_0^{(2)}(\cdot)$ denotes the smooth baseline hazard as a function of time. The effect of the covariate x_5 in the copula parameter $\vartheta^{(c)}$ leads to various degrees of dependence in either positive or negative directions. We consider cases where the the non-time-to-event margin is set as a binary random variable, resembling the data analysed in Section 4 in the main manuscript. The true baseline functions (hazard, cumulative hazard and survival) used in our simulations are shown in Figure D1(a), (b) and (c), respectively. The functions plotted using red lines correspond to the baseline functions used for simulation scenarios with heavy censoring rates. The hazard rate of the event times used in mild censoring scenarios achieves its maximum at time point $time_{mild}^{max\ haz} = 1.809$, whereas in the heavy censoring scenarios the mode of the hazard rate is at time point $time_{heavy}^{max\ haz} = 3.417$, see Figure D1(a).

We consider the sample sizes $n_1 = 750$, and $n_2 = 1500$. Censoring times are drawn from independent continuous uniform distributions after generating the true event times. The support of the distributions of the censoring times are set such that the censoring rates are approximately 20% (mild censoring case) and 70% (heavy censoring case), respectively.

Bivariate mixed outcomes are generated using the Gaussian, Frank, Clayton and Gumbel copulas. Since the Clayton and Gumbel copulas can only model positive dependence, we employ 90° rotations to generate and fit negative dependence between the observations using these two copulas. We first sample the covariates from independent uniform distributions with

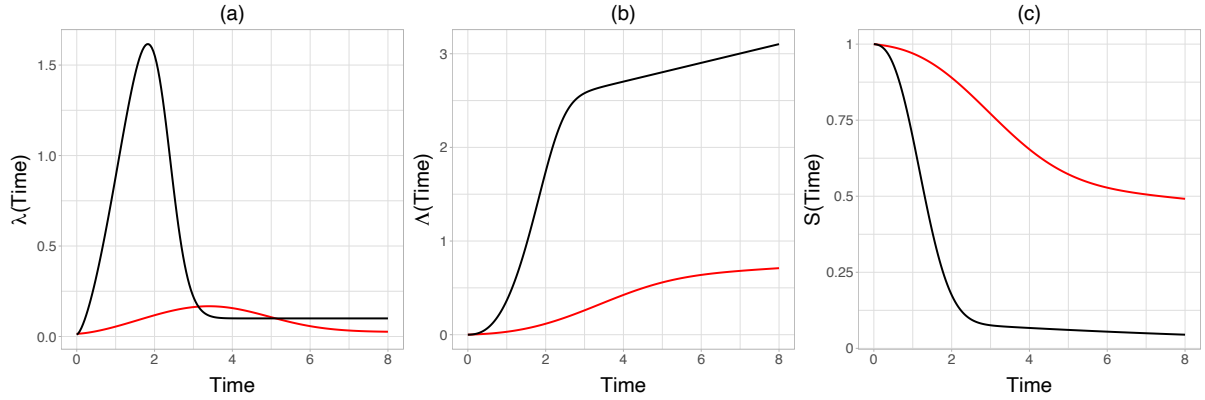

Figure D1: True baseline functions (hazard rate (a), cumulative hazard (b) and survival function (c)) used to generate synthetic data. Black lines correspond to mild censoring, whereas red lines were used for heavy censoring scenarios.

support in the interval $[-1, +1]$. Then, we construct the additive predictor for the dependence parameter $\eta^{(c)}$ and sample bivariate uniform margins from a copula with dependence parameter $\vartheta^{(c)} = g^{-1}(\eta^{(c)})$ using the `R` package `VineCopula` Nagler et al. (2022). See Table D3 for examples of suitable link functions. Afterwards we compute the distribution parameters of the margins and obtain the non-time-to-event response by applying the quantile function of its corresponding distribution to the first sampled uniform margin from the copula. The generation of our time-to-event margin is based on Marra and Radice (2019). As a starting point, we construct the survival function using the following predictor:

$$\eta^{(2)} = \log(-\log(s_0^{(2)}(t))) + \beta_1^{(2)} x_2 + \beta_2^{(2)} x_4 \Leftrightarrow S(t) = \exp\left(-\exp\left(\eta^{(2))}\right)\right),$$

where $s_0^{(2)}(\cdot)$ is the baseline survival function. Then, we numerically invert the second uniform margin sampled from the copula together with the survival function in order to obtain the true event times T. The censoring times T^{cens} are obtained by sampling from a uniform distribution. Finally, the observed time-to-event responses are created using $\tilde{T} = \min\{T, T^{cens}\}$ and $\delta = \{T \leq T^{cens}\}$.

In total, $R = 100$ Monte Carlo replications are used throughout this section. The number of discrete time intervals is set to $J = 20$ throughout the simulation studies presented in this section, although we remark that setting the number of intervals J to a larger value or to the number of unique event times (e.g. in the PW) case would result in a finer time grid to estimate the hazard rate from. We remark that we do not treat the number of intervals as a tuning parameter / hyperparameter in our approach since the regularization induced via the quadratic penalty addresses the compromise between overfitting data and smoothness of the estimated baseline hazard. Table D1 depicts the scenarios considered in our simulation studies. The `R` code used to replicate the simulations can also be found in the following link: `https://github.com/GuilleBriseno/DTPW_DistCopulaReg`.

Table D1: Overview of scenarios considered in the simulation studies. All fitted using Gaussian, Frank, Clayton, Clayton 90°, Gumbel and Gumbel 90° copulas.

Scenario	Non-time-to-event	Time-to-event	Censoring rate	n
1	Binary	*DT*	Mild ($\approx 20\%$)	750
2	Binary	*DT*	Mild ($\approx 20\%$)	1500
3	Binary	*DT*	Heavy ($\approx 70\%$)	750
4	Binary	*DT*	Heavy ($\approx 70\%$)	1500
5	Binary	*PW*	Mild ($\approx 20\%$)	750
6	Binary	*PW*	Mild ($\approx 20\%$)	1500
7	Binary	*PW*	Heavy ($\approx 70\%$)	750
8	Binary	*PW*	Heavy ($\approx 70\%$)	1500

Results for Q1

The estimated baseline hazard rates using the *DT* approach are show in Figure D2. Under the chosen configuration for the number of intervals ($J = 20$), the *DT* approach is able to estimate the underlying baseline hazard rather well. For both mild and heavy censoring cases, it can be seen that the estimated baseline hazard tends to exhibit excessive "wiggliness" towards the end of the study time period, with heavy censoring scenarios showing the most variability in the aforementioned time range. This is expected, since estimation under these circumstances is rather challenging. The additional variability under both censoring regimes is also due to the low number of observations that are observed in that time range. Overall, increasing the sample size leads to a reduced variability in the estimated functions, this can be seen in the blue lines ($n = 1500$) being surrounded by red lines ($n = 750$) in all panels of Figure D2. The *DT* approach is able to capture the most important aspects of the underlying true baseline hazard rate, these would be for us its shape and (approximate) time point of its mode, see the vertical dotted red lines in the aforementioned figures.

Figures D3 shows the estimated survival functions using the *DT* approach. Compared to the baseline hazard rate, the estimated baseline survival function obtained from the *DT* approach exhibits a much more robust behaviour. As expected, increasing the sample size leads to a reduced variability in the estimated functions, this reduction can be better appreciated in the panels corresponding to heavy censoring scenarios in Figure D3. Towards the end of the study time, the fits start to exhibit more variability, especially in cases where heavy censoring is present. As previously mentioned, this behaviour is expected due to a low number of observations in that time range, as well as the high censoring prevalence. Similar to the baseline hazard rate, the estimated baseline survival functions exhibit the same behaviour.

Table D2: Definition of the considered copulas. $F_\bullet$ with $\bullet \in \{1, 2\}$ denotes the CDF of the margins. Rotations by 90° are obtained using: $C_{90} = F_2 - C(1 - F_1, F_2)$. The term $D_1(\vartheta^{(c)}) = \int_0^{\vartheta^{(c)}} \frac{t}{\exp(t)-1} dt$ is the Debye function and $\Phi_2(\cdot, \cdot)$ denotes the CDF of the bivariate Gaussian distribution with correlation coefficient $\vartheta^{(c)}$.

Copula	$C(F_1, F_2; \vartheta^{(c)})$	Range of $\vartheta^{(c)}$	Kendall's τ
Gauss	$\Phi_2(\Phi^{-1}(F_1), \Phi^{-1}(F_2); \vartheta^{(c)})$	$\vartheta^{(c)} \in [-1, 1]$	$\frac{2}{\pi} \arcsin(\vartheta^{(c)})$
Frank	$-\vartheta^{(c)^{-1}} \log\Big(1 + (\exp(-\vartheta^{(c)} F_1) - 1)$ $(\exp(-\vartheta^{(c)} F_2) - 1)/(\exp(-\vartheta^{(c)}) - 1)\Big)$	$\vartheta^{(c)} \in \mathbb{R} \setminus \{0\}$	$1 - \frac{4}{\vartheta^{(c)}}[1 - D_1(\vartheta^{(c)})]$
Clayton	$(F_1^{-\vartheta^{(c)}} + F_2^{-\vartheta^{(c)}} - 1)^{-1/\vartheta^{(c)}}$	$\vartheta^{(c)} \in (0, \infty)$	$\frac{\vartheta^{(c)}}{\vartheta^{(c)}+2}$
Gumbel	$\exp\left[-\left\{(-\log(F_1))^{\vartheta^{(c)}} + (-\log(F_2))^{\vartheta^{(c)}}\right\}^{1/\vartheta^{(c)}}\right]$	$\vartheta^{(c)} \in [1, \infty)$	$1 - \frac{1}{\vartheta^{(c)}}$

As shown in Figure D4, estimating the time-to-event margin using the *PW* approach produces results similar to those of the *DT* method. Overall, it can be seen that the *PW* approach correctly captures the behaviour of the underlying hazard rate. The estimated hazard rate reaches its mode close to the time point of the true mode. Increasing the sample size reduces the variability in the estimates, as seen from the blue lines being enclosed by the red lines.

Results for Q2.

Figure D6 shows the results of the simulation studies for the *DT* approach. It can be seen that in general the proposed approach is able to recover the true coefficient values that have an effect on either the respective margins or the dependence parameter with satisfactory performance. The increased variation of the estimated coefficients in the copula dependence parameter increases in cases with heavy censoring, but this behaviour is expected in such challenging scenarios. Estimating the coefficients in the dependence parameter seems to be more challenging when using non-elliptical / non-symmetric copulas in cases of negative dependence (i.e. rotations of the Clayton and Gumbel copulas). The *PW* approach produces results very similar to those obtained using the *DT* counterpart. Figure D7 displays the bias of the model coefficients across all distribution parameters and fitted copulas. The coefficients of all distribution parameters are estimated correctly using elliptical copulas (Gauss and Frank) as well as non-rotated Gumbel and Clayton copulas. Estimation of the effects on the dependence parameter appears to be more challenging for 90 degree rotations of the aforementioned copulas, in particular for the Clayton copula.

Table D3: Link and response functions for a generic parameter ϑ. $\Phi(\cdot)$ denotes the CDF of the standard Gaussian distribution and $\Phi^{-1}(\cdot)$ its quantile function.

Range	Response function: $g^{-1}(\eta) = \vartheta$	Link function: $g(\vartheta) = \eta$
$\vartheta \in [0, 1]$	$\frac{\exp(\eta)}{1+\exp(\eta)}$	$\ln\left(\frac{\vartheta}{1-\vartheta}\right)$
	$\Phi(\eta)$	$\Phi^{-1}(\vartheta)$
	$1 - \exp(-\exp(\eta))$	$\ln(-\ln(1-\vartheta))$
	$\exp(-\exp(-\eta))$	$-\ln(-\ln(\vartheta))$
	$\exp(\eta)$	$\ln(\vartheta)$
$\vartheta \geq 0$	$\exp(\eta)$	$\ln(\vartheta)$
$\vartheta \in [1, \infty)$	$\exp(\eta) + 1$	$\ln(\vartheta - 1)$
$\vartheta \in [-1,\ 1]$	$\tanh(\eta)$	$\tanh^{-1}(\vartheta)$

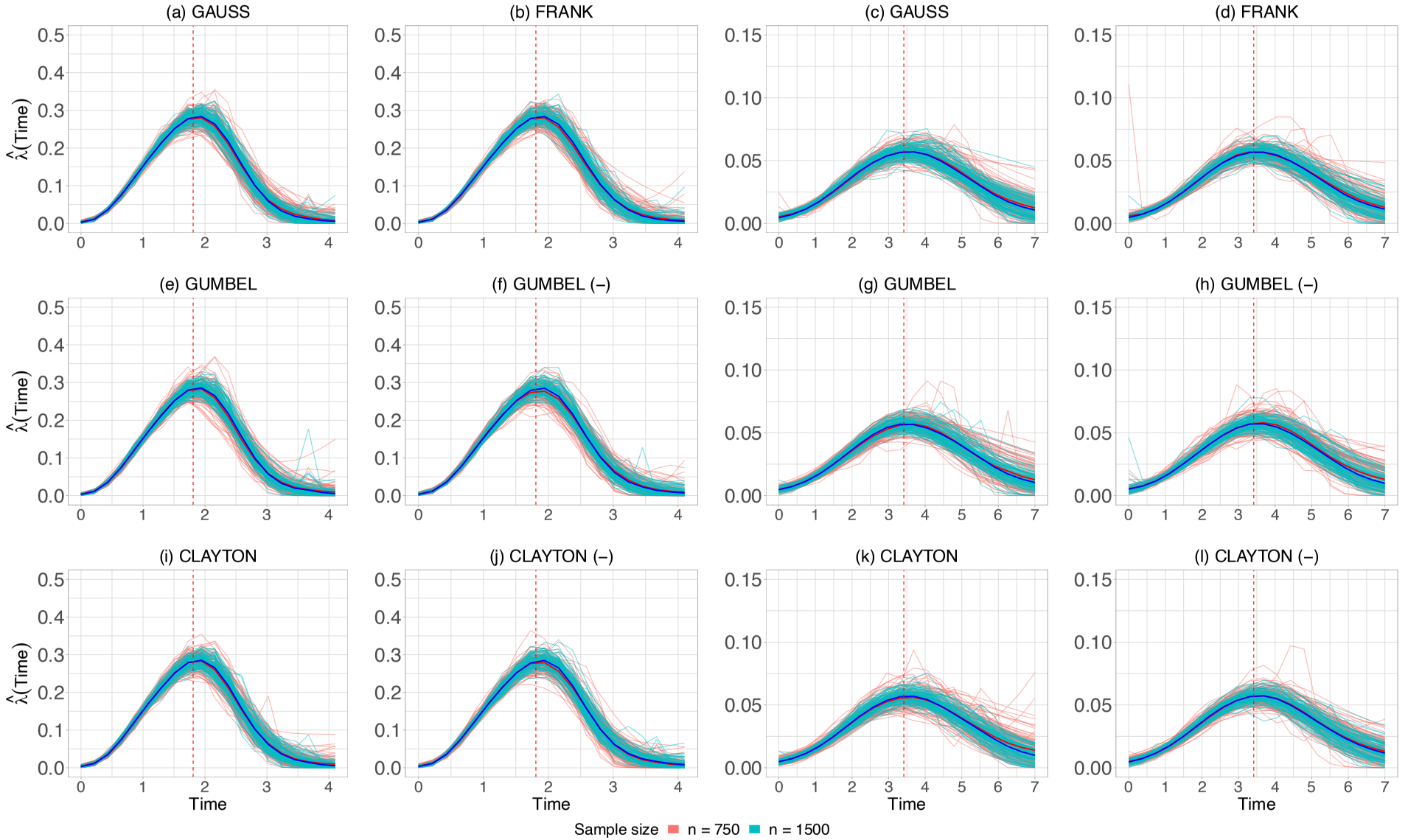

Figure D2: Estimated baseline hazard rate using the *DT* approach across sample sizes and copula functions given a binary non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $time_{mild}^{max\ haz} = 1.809$ and $time_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

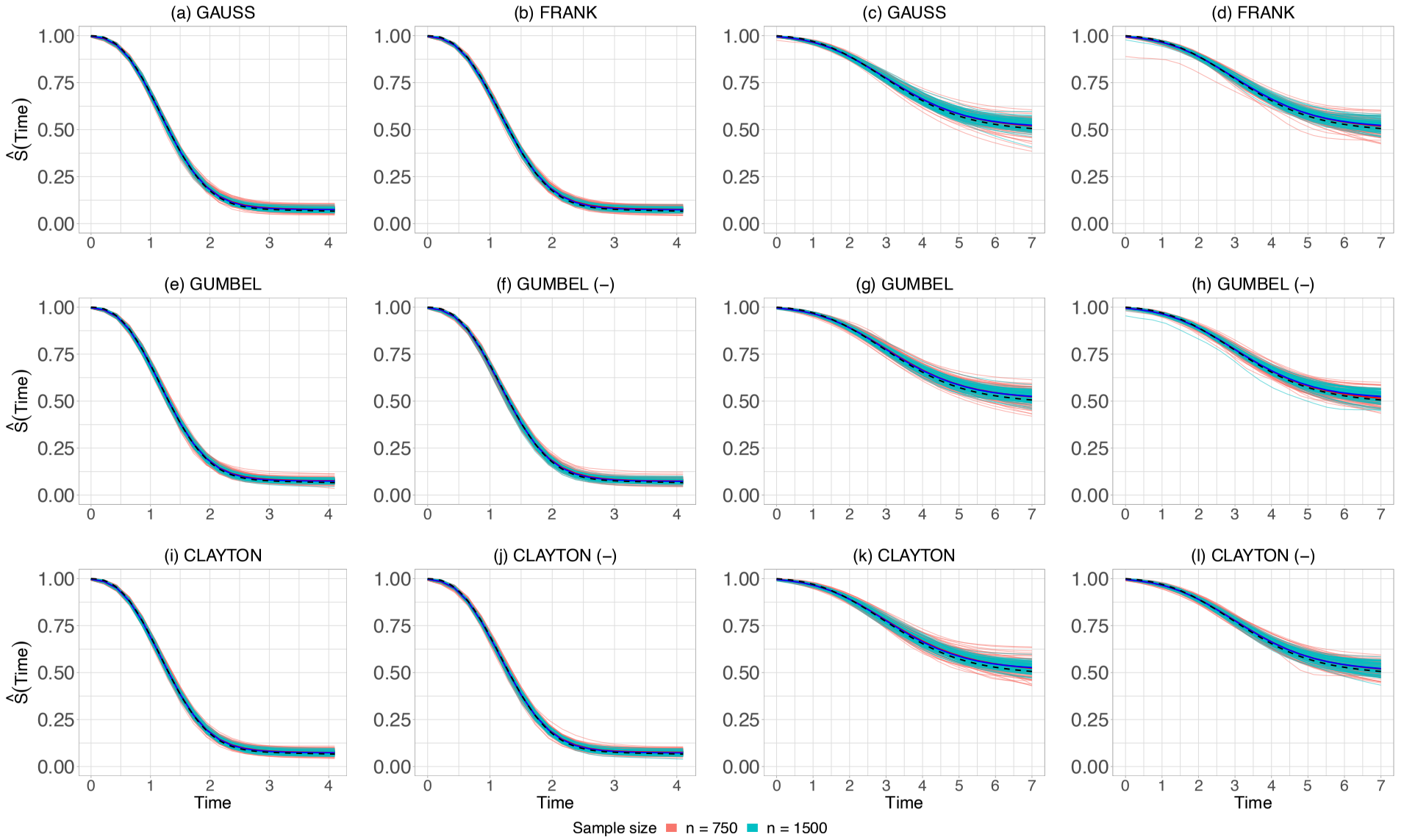

Figure D3: Estimated baseline survival function using the *DT* approach across sample sizes and copula functions given a binary non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

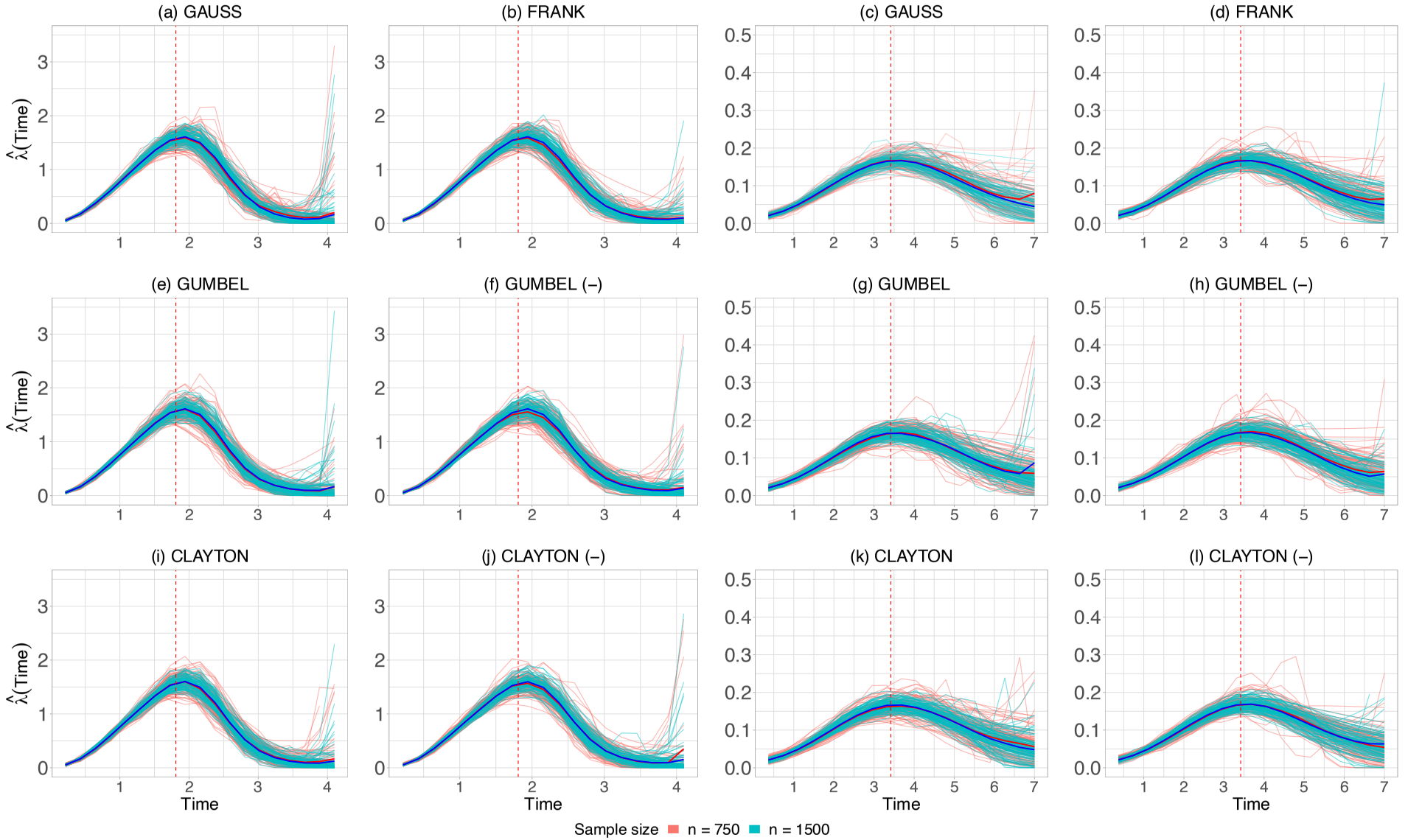

Figure D4: Estimated baseline hazard rate using the PW approach across sample sizes and copula functions given a binary non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $time_{mild}^{max\ haz} = 1.809$ and $time_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

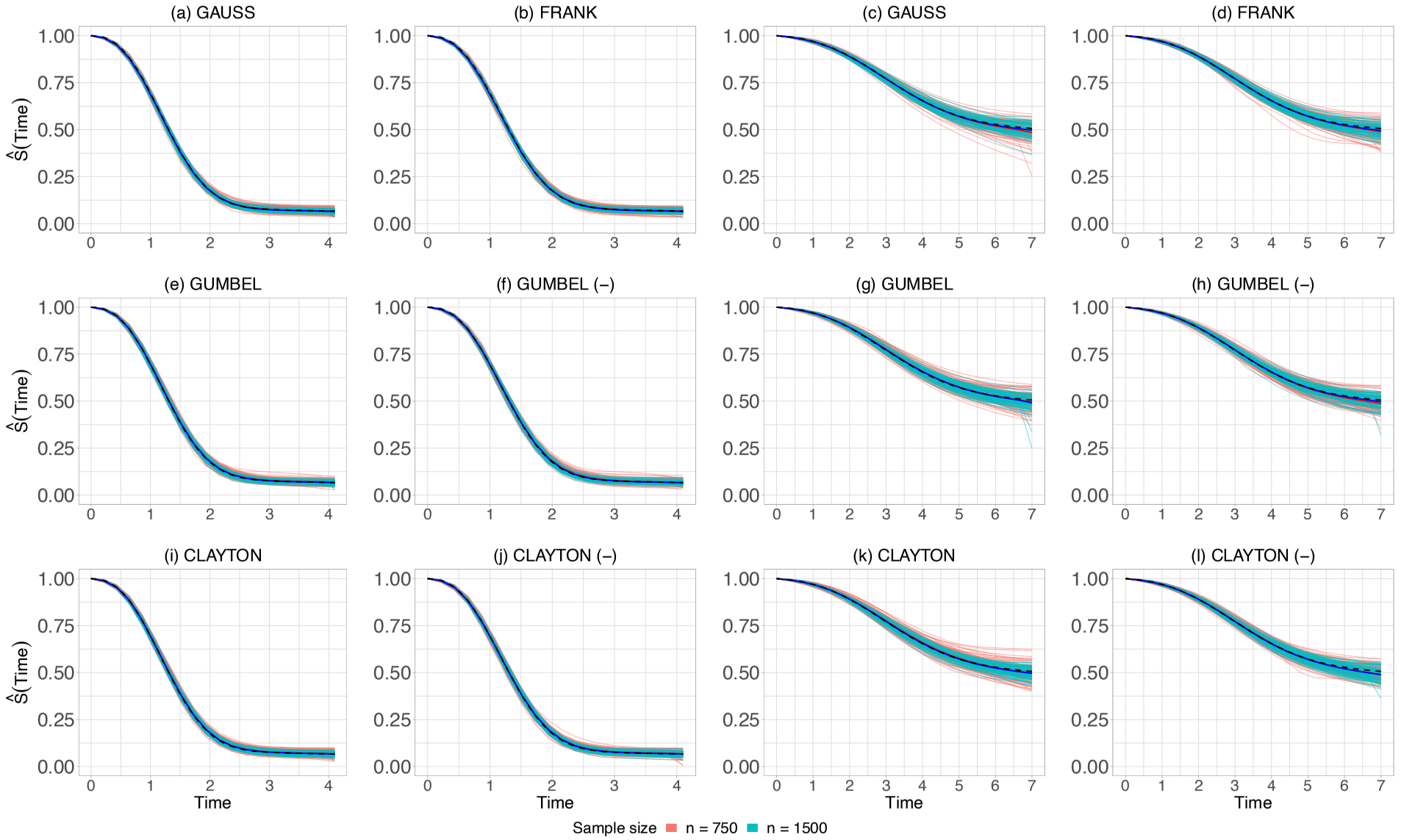

Figure D5: Estimated baseline survival function using the PW approach across sample sizes and copula functions given a binary non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

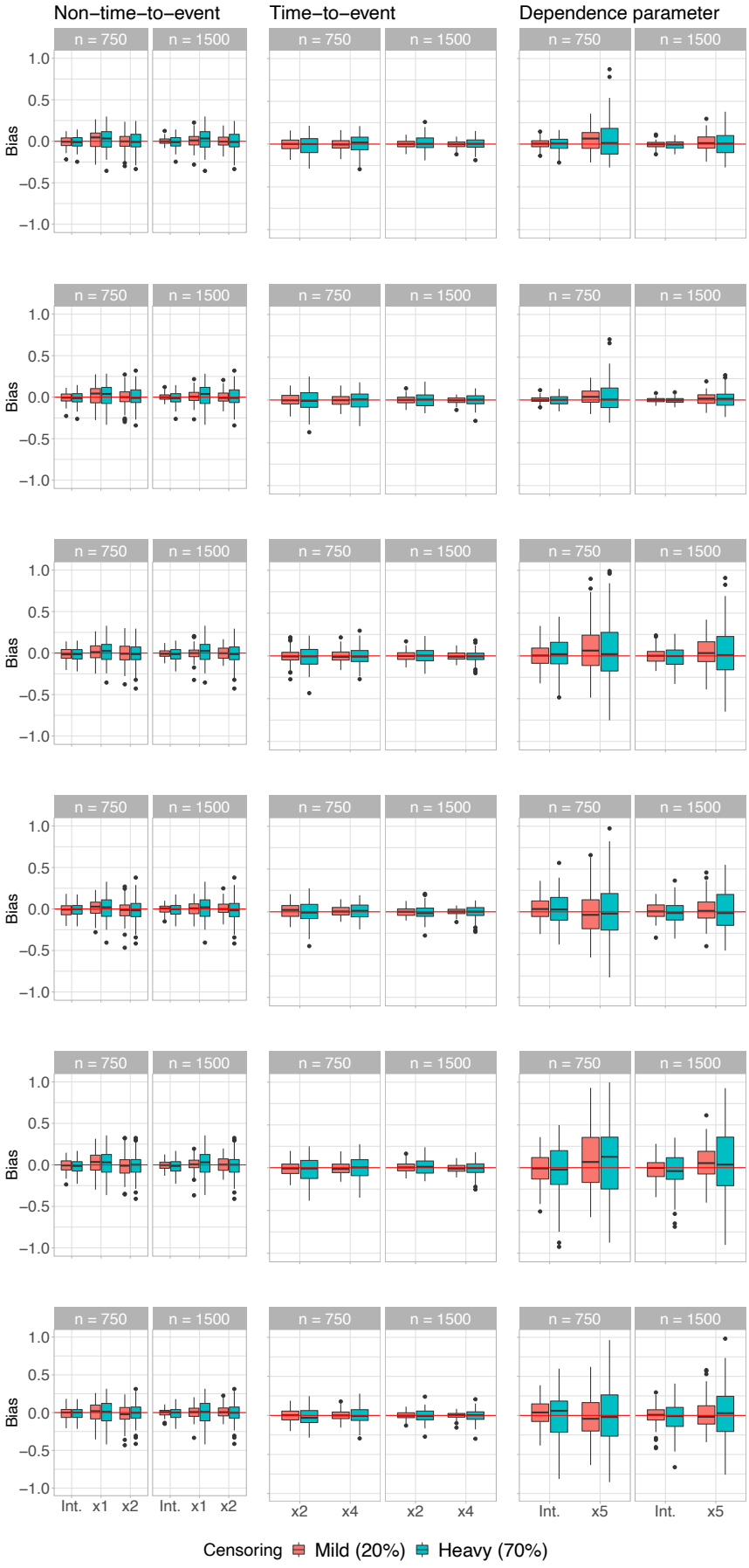

Figure D6: Bias of coefficients using the *DT* approach. Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.

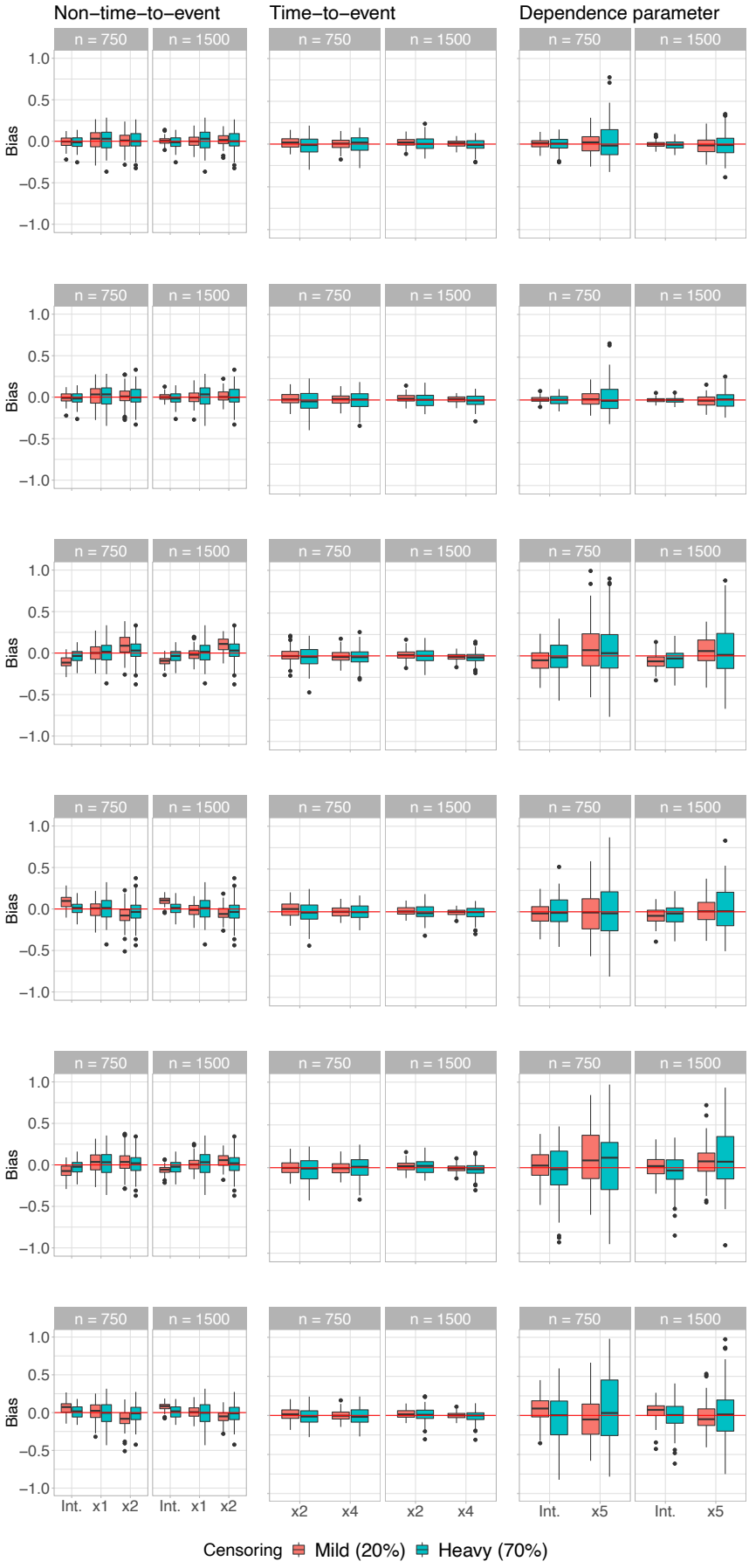

Figure D7: Bias of coefficients using the *PW* approach. Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.

Appendix C: Boosting distributional copula regression for bivariate binary, discrete and mixed responses (with Supplement)

Joint work with Nadja Klein, Hannah Klinkhammer, and Andreas Mayr.

Briseño Sanchez, G., Klein, N., Klinkhammer, H., & Mayr, A. (2025). Boosting distributional copula regression for bivariate binary, discrete and mixed responses. *Statistical Methods in Medical Research*, https://doi.org/10.1177/09622802241313294.

Published in *Statistical Methods in Medical Research.* https://doi.org/10.1177/09622802241313294.

Original Research Article

SMMR
STATISTICAL METHODS IN MEDICAL RESEARCH

Boosting distributional copula regression for bivariate binary, discrete and mixed responses

Statistical Methods in Medical Research
1–16


Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/09622802241313294
journals.sagepub.com/home/smm


Guillermo Briseño Sanchez[1], Nadja Klein[1], Hannah Klinkhammer[2] and Andreas Mayr[2]

Abstract

Motivated by challenges in the analysis of biomedical data and observational studies, we develop statistical boosting for the general class of bivariate distributional copula regression with arbitrary marginal distributions, which is suited for binary, count, continuous or mixed outcomes. To arrive at a flexible model for the entire conditional distribution, not only the marginal distribution parameters but also the copula parameters are related to covariates through additive predictors. We suggest estimation by means of an adapted component-wise gradient boosting algorithm. A key benefit of boosting as opposed to classical likelihood or Bayesian estimation is the implicit data-driven variable selection mechanism as well as shrinkage. To the best of our knowledge, our implementation is the only one that combines a wide range of covariate effects, marginal distributions, copula functions, and implicit data-driven variable selection. We showcase the versatility of our approach to data from genetic epidemiology, healthcare utilization and childhood undernutrition. Our developments are implemented in the R package `gamboostLSS`, fostering transparent and reproducible research.

Keywords

Dependence modelling, GAMLSS, model-based boosting, shrinkage, variable selection

1 Introduction

Distributional regression models have gained considerable prominence in statistical research over the last decade, thereby moving the focus from modelling the conditional mean of the response variable (as done in classical regression) towards modelling the entire conditional distribution.[1] Such models capable of describing the complete distribution are highly relevant in biomedical research, as they allow to explore variables that impact not only the average value of biomarkers, phenotypes or scores but also other quantities such as variance or quantiles. Common examples are the construction of reference curves or growth charts, where skewness is often covariate-specific[2,3]; or bivariate time-to-event data.[4]

Several distinct approaches to distributional regression for univariate responses exist, see Klein[1] for a recent review. Our framework builds on generalized additive models for location, scale shape (GAMLSS),[5] which allow us to relate all distribution parameters of an arbitrary univariate parametric distribution to covariates. A simple example is a GAMLSS for the Gaussian regression model, in which not only the expectation, but also the standard deviation can be related to covariates. This allows, for example, model heteroscedasticity. While originally proposed for univariate responses, GAMLSS has been extended to accommodate regression models for multivariate responses,[6] although practically most existing approaches are

[1]Methods for Big Data, Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany
[2]Department of Medical Biometrics, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany

Corresponding author:
Guillermo Briseño Sanchez, Methods for Big Data, Scientific Computing Center, Karlsruhe Institute of Technology, Zirkel 2, 76131 Karlsruhe, Germany.
Email: guillermo.briseno-sanchez@kit.edu

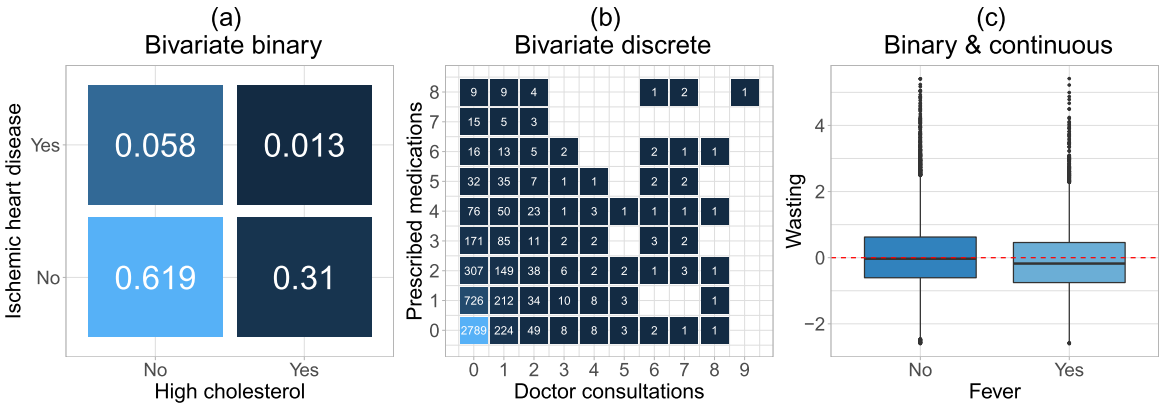

Figure 1. Responses in our applications analysed in Section 4: (a) binary–binary response (numbers indicate proportions): high cholesterol and chronic ischemic heart disease; (b) count–count response (numbers indicate cases): doctor visits and medical prescriptions; and (c) binary–continuous response: fever and wasting (indicator for acute undernutrition).

limited to the bivariate case.[7–9] While parametric bivariate distributions such as the bivariate Gaussian, bivariate Bernoulli or bivariate Poisson offer an avenue for modelling bivariate responses, they also impose limitations on the distribution of the margins, for example, being univariate Gaussian or Poisson. A flexible alternative way to construct bivariate distributions is copulas.[10] This approach allows to linking of arbitrary marginal distributions through a copula function, reflecting the association between the components. The literature on copula modelling, including the regression setting, is vast see, for example , Smith[11] for a review.

Reflecting the diversity of response types in our biomedical applications, in this paper, we are particularly concerned with situations where the response variable is a bivariate vector $\boldsymbol{Y} = (Y_1, Y_2)^\top$ with dependent components on possibly different domains. We construct bivariate distributions for such situations via conditional copulas and parametric margins; and allow all distributional parameters of the joint density to depend on covariates. Estimation is realized jointly rather than employing a two-step procedure frequently employed in copula models. Recent contributions that are akin to ours can be found in Marra and Radice[12] featuring a bivariate continuous response, Marra and Radice[13] using bivariate binary outcomes, van der Wurp et al.[14] studying bivariate count responses, as well as Klein et al.[15] analysing a mixed binary and continuous response. All these contributions showed how to construct highly flexible bivariate copula regression models that are able to accommodate a wide range of covariate effects as well as response types. However, the substantial flexibility inherent in this model class of distributional copula regression models notably exacerbates the issue of variable selection—a challenge that currently remains unaddressed within the specific models we are considering.

Our methodological contribution builds on the recent work by Hans et al.,[16] who estimated bivariate distributional copula regression models via a component-wise gradient boosting framework. However, in this approach, the response variables are both required to be strictly continuous. In many biomedical applications (but not only there), data are often recorded at a discretized scale (e.g. symptoms present yes/no) or the responses of interest actually depict a phenomenon expressed through discrete numbers/positive integers as in, for example, the number of doctor appointments or the number of prescription medications designated to a patient. At the time of writing (December 2023), a search in PubMed (https://pubmed.ncbi.nlm.nih.gov) returns 395,078 and 24,439 results for "logistic regression" and "Poisson regression" since 2010, respectively, highlighting the prevalence of binary and count responses. It may also be the case that the biomedical outcome is expressed as a combination of responses that lie in different domains, for example, a binary indicator and a continuous measurement reflecting a disease (or symptom) indicator and an undernutrition score. The three aforementioned types of responses are the ones we consider later in Section 4 and the marginal distributions are visualized in Figure 1.

Recent work by Strömer et al.[17] combined multivariate distributional regression with gradient boosting in order to fit interpretable and highly flexible regression models in high-dimensional biomedical settings for bivariate continuous, bivariate binary and bivariate count responses. Their work considered the bivariate Poisson and the bivariate Bernoulli distributions, which suffer from some limitations. On the one hand, the bivariate Poisson distribution is only able to model positive association structures between the margins. On the other hand, the bivariate Bernoulli distribution models the association between the marginal responses by means of the 'odds ratio', whose ease of interpretation remains at best questioned, see, for example, Norton et al.[18] Furthermore, the marginal distributions of the components in the response vector are assumed to be of the same type, that is, the margins of a bivariate Bernoulli/Poisson distribution must be univariate Bernoulli/Poisson distributions. Such a restrictive assumption might not always be supported by the data. For example,

in one of our applications where we study childhood undernutrition via the joint distribution of *wasting*, a continuous indicator for acute undernutrition as reflected by low weight-for-height (in comparison to a reference population), and a binary indicator for fever within the two weeks preceding a survey interview.

The aim of this article is threefold: First, we build upon Hans et al.[16] and extend the class of boosting bivariate distributional copula regression models to arbitrary margins on different domains. Second, we expand the catalogue of copula functions and families of marginal distributions available for the publicly available R[19] package `gamboostLSS`.[20] These new additions allow for conducting data-driven variable selection and shrinkage in both low- and high-dimensional applications, where the number of candidate variables (p) may greatly exceed the number of observations (n). Third, we demonstrate the versatility and wide applicability of our approach through three biomedical applications.

The rest of the article is structured as follows: Section 2 reviews distributional copula regression with different types of responses and outlines our boosting algorithm. Section 3 summarizes our simulation studies as well as their respective results. Section 4 presents the three case studies where we analyse data from epidemiological applications in genetic epidemiology, healthcare and public health in developing countries. We additionally illustrate the model-building process that involves selecting marginal distributions as well as copula distributions. Lastly, a discussion is given in Section 5. Supplemental material C contains further details on the simulation studies.

2 Bivariate distributional copula regression

2.1 Model structure

We assume that the ith observation of a response, with $i = 1, \ldots, n$, follows a parametric distribution. In the context of bivariate responses considered here, the joint distribution of the random vector $\boldsymbol{Y} = (Y_{i1}, Y_{i2})^\top$ is denoted by $P(Y_1 \leq y_{i1},\ Y_2 \leq y_{i2}; \boldsymbol{\vartheta}_i) = F_{1,2}(y_{i1}, y_{i2}; \boldsymbol{\vartheta}_i)$, where $F_{1,2}(\cdot; \boldsymbol{\vartheta}_i)$ represents the joint cumulative distribution function (CDF) parameterized through a K-dimensional parameter vector $\boldsymbol{\vartheta}_i = (\vartheta_{i1}, \ldots, \vartheta_{iK})^\top$. Rather than assuming a joint parametric distribution for $\boldsymbol{Y}$, we resort to a copula-based approach using Sklar's theorem.[10] This theorem states that any bivariate distribution can be written as

$$F(y_{i1}, y_{i2}; \boldsymbol{\vartheta}_i) = C[F_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}); \boldsymbol{\vartheta}_i^{(c)}] \tag{1}$$

where $C(\cdot, \cdot) : [0, 1]^2 \to [0, 1]$ is the CDF of a bivariate parametric copula function with parameters $\boldsymbol{\vartheta}_i^{(c)} \in \mathbb{R}^{K_c}$. The copula *links* the possibly different parametric marginal distributions with CDFs F_1, F_2 and respective parameter vectors $\boldsymbol{\vartheta}_i^{(1)} \in \mathbb{R}^{K_1}$, $\boldsymbol{\vartheta}_i^{(2)} \in \mathbb{R}^{K_2}$ to arrive at the joint bivariate distribution. In what follows, we consider one-parametric bivariate copulas and refer to $\boldsymbol{\vartheta}_i^{(c)} = \vartheta_i^{(c)}$ as the corresponding scalar association parameter that determines the strength of the association between the marginal responses. Table SA1 in Supplemental material A details the implemented copulas in the R add-on package `gamboostLSS`.

Let now $K = K_1 + K_2 + K_c = K_1 + K_2 + 1$ denote the total number of distribution parameters in the bivariate distribution and $\boldsymbol{\vartheta}_i^{(1)} = (\vartheta_{i1}^{(1)}, \ldots, \vartheta_{iK_1}^{(1)})^\top$, $\boldsymbol{\vartheta}_i^{(2)} = (\vartheta_{i1}^{(2)}, \ldots, \vartheta_{iK_2}^{(2)})^\top$, be the vectors containing all parameters that correspond to the respective marginal distributions. All K parameters of the bivariate distribution are then stored in the vector $\boldsymbol{\vartheta}_i = ((\boldsymbol{\vartheta}_i^{(1)})^\top, (\boldsymbol{\vartheta}_i^{(2)})^\top, \boldsymbol{\vartheta}_i^{(c)})^\top$. The distributional copula regression approach relates each component of $\boldsymbol{\vartheta}_i$ to possibly different subvectors of the covariate information $\boldsymbol{x}_i$. More precisely, we employ structured additive predictors of the form

$$g_k^{(\bullet)}(\vartheta_{ik}^{(\bullet)}) = \eta_{ik}^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} s_{rk}^{(\bullet)}(x_{ir}) \tag{2}$$

where $g_k^{(\bullet)}(\cdot)$ are suitable link functions with corresponding inverse functions $h_k^{(\bullet)}(\cdot)$ that ensure potential parameter space restrictions. The symbol $\bullet \in \{1, 2, c\}$ and the summation limit $P_k^{(\bullet)}$ emphasize that the individual parameters $\vartheta_{ik}^{(\bullet)}$ do not necessarily have to be modelled using the same subset of covariates. The coefficients $\beta_{0k}^{(\bullet)}$ are parameter-specific intercepts and $s_{rk}^{(\bullet)}(\cdot)$ are smooth functions that can accommodate a wide range of functional forms of the covariates, such as linear, non-linear, or spatial effects. Each covariate effect is modelled by appropriate basis function expansions of the form

$$s_{rk}^{(\bullet)}(x) = \sum_{l=1}^{L_{rk}^{(\bullet)}} \beta_{rk,l}^{(\bullet)} B_{rk,l}^{(\bullet)}(x)$$

where $B^{(\bullet)}_{rk,l}(x)$ is a suitable basis function evaluated at the observed covariate value and $\beta^{(\bullet)}_{rk,l}$ are generic coefficients to be estimated. As a final remark, Sklar's theorem guarantees that the copula characterising the joint distribution of $\mathbf{Y}_i$ is unique only if the marginal responses are continuous. When discrete margins are present, the copula is uniquely defined only on the range of the marginal CDFs. However, within our framework, identifiability should not be an issue for two reasons. First, we fix the parametric form of the joint distribution a priori by making choices for the marginal distributions and the copula function. Thereby, potential identifiability issues that arise when one is interested in learning, for example, the copula and the marginals in a nonparametric framework without an a priori fixed structure, are not present. Second, identifiability is ensured in our regression setting where all parameters of the distribution are observation-specific. For this, consider two observations in the sample, say i and i', with the same observed marginal response ($y_{ij} = y_{ij'}$) but different covariate values for $j = 1, 2$. Modelling the parameters of the respective marginal distributions as functions of covariates results in different estimates for the marginal CDFs, that is, $\hat{F}_j(y_{ij}; \hat{\boldsymbol{\vartheta}}_i) \neq \hat{F}_j(y_{ij'}; \hat{\boldsymbol{\vartheta}}_{i'}), j = 1, 2$. Hence a richer range of the estimated CDFs of the discrete marginal distributions is obtained, mitigating the identification issue when using these types of marginal responses. This has been also pointed out by other researchers.[21–24]

2.2 Relevant examples of bivariate responses

In the following, we briefly describe the bivariate responses relevant to our applications. The respective choices of corresponding marginal distributions are summarized in Table SA2 in Supplemental material A together with main characteristics, such as expectation and variance.

2.2.1 Bivariate binary responses

We begin by considering the case $Y_{ij} \in \{0, 1\}, j = 1, 2$. The individual marginal probabilities of observing $y_{ij} = 1$ are modelled via $P(Y_{ij} = 1; \vartheta_i^{(j)}) = \vartheta_i^{(j)} = h^{(j)}(\eta_i^{(j)}) =: p_i^{1(j)}, j = 1, 2$, where the response function can be any function suitable for parameters whose range is the unit interval $[0, 1]$, for example, logit, probit and cloglog link functions. Then,

$$\begin{aligned} P(Y_{i1} = 1, Y_{i2} = 1; \boldsymbol{\vartheta}_i) &= C[P(Y_{i1} = 1; \vartheta_i^{(1)}), P(Y_{i2} = 1; \vartheta_i^{(2)}); \vartheta_i^{(c)}] \\ &=: p_i^{11} \end{aligned}$$

The joint probability mass function consists of the four possible outcomes of the binary responses, that is, $(y_{i1}, y_{i2}) \in \{(1, 1), (1, 0), (0, 1), (0, 0)\}$. This leads to the following log-likelihood contribution of the ith observation.

$$\begin{aligned} \ell_i = {} & y_{i1}y_{i2} \log(p_i^{11}) + y_{i1}(1 - y_{i2}) \log(p_i^{1(1)} - p_i^{11}) \\ & + (1 - y_{i1})y_{i2} \log(p_i^{1(2)} - p_i^{11}) \\ & + (1 - y_{i1})(1 - y_{i2}) \log(1 - p_i^{1(1)} - p_i^{1(2)} + p_i^{11}) \end{aligned} \tag{3}$$

Note that our implementation allows the individual marginal probabilities to be modelled using different link functions.

2.2.2 Bivariate discrete responses

Each marginal response is a count variable, that is, $Y_{ij} \in \mathbb{N}_{\geq 0}, j = 1, 2$. Here we denote with $P(Y_{ij} \leq y_{ij}; \boldsymbol{\vartheta}_i^{(j)}) = F_j(y_{ij}; \boldsymbol{\vartheta}_i^{(j)})$ the marginal CDFs, and with $P(Y_{ij} = y_{ij}; \boldsymbol{\vartheta}_i^{(j)}) = f_j(y_{ij}; \boldsymbol{\vartheta}_i^{(j)})$ the marginal probability distribution functions of Y_{ij}. Similar to van der Wurp et al.,[14] we compute $P(Y_{ij} = y_{ij} - 1; \boldsymbol{\vartheta}_i^{(j)}) = F_j(y_{ij}; \boldsymbol{\vartheta}_i^{(j)}) - f_j(y_{ij}; \boldsymbol{\vartheta}_i^{(j)})$ in order to avoid a (trivial) evaluation of the CDF of Y_{ij} with a negative argument in case that $y_{ij} = 0, j = 1, 2$. The log-likelihood function of the ith observation is then given by

$$\begin{aligned} \ell_i = \log \Big\{ & C[F_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}); \vartheta_i^{(c)}] \\ - \; & C[F_1(y_{i1};\ \boldsymbol{\vartheta}_i^{(1)}) - f_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)};\ \vartheta_i^{(c)})] \\ - \; & C[F_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}) - f_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}] \\ + \; & C[F_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}) - f_1(y_{i1}; \boldsymbol{\vartheta}_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}) - f_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}] \Big\} \end{aligned} \tag{4}$$

We have implemented various discrete distributions, including the Poisson and Geometric distributions. Additionally, we have integrated two-parameter count distributions designed for over-dispersed data such as the negative binomial (Type I).

To handle count data characterized by an excess of zero observations, we have included zero-inflated and zero-altered distributions. These include models such as the zero-altered logarithmic, zero-altered negative binomial, zero-inflated Poisson and zero-inflated negative binomial distributions, see Table SA2 in Supplemental material A for a detailed description as well as Rigby et al.

2.2.3 Bivariate mixed binary–continuous responses

When one response component is continuous and the other binary, we follow Klein et al.[15] and resort to a latent variable representation of the regression model for the binary component. Without loss of generality, let the first component of the bivariate vector be the binary variable, that is, $Y_{i1} \in \{0, 1\}$. The binary response Y_{i1} is then determined by an unobserved, latent variable Y_{i1}^* with parametric CDF $F_1^*(y_{i1}^*; \boldsymbol{\vartheta}_i^{(1)})$ through the mechanism: $Y_{i1} = \mathbb{1}(Y_{i1}^* > 0)$, where $\mathbb{1}(\cdot)$ is the indicator function. Then it follows that $P(Y_{i1} = 0; \vartheta_i^{(1)}) = F_1(0; \vartheta_i^{(1)}) = F_1^*(0; \vartheta_i^{(1)}) = P(Y_{i1}^* \leq 0; \vartheta_i^{(1)})$, in other words, the CDFs of the binary and latent variables coincide at $y_{i1} = y_{i1}^* = 0$. With this representation, the joint bivariate distribution is

$$\begin{aligned} P(Y_{i1} = 0, Y_{i2} \leq y_{i2}; \boldsymbol{\vartheta}_i) &= P(Y_{i1}^* \leq 0, Y_{i2} \leq y_{i2}) \\ &= C[F_1^*(0; \vartheta_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)});\ \vartheta_i^{(c)}] \end{aligned}$$

from which we obtain the log-likelihood contribution:

$$\begin{aligned} \ell_i = (1 - y_{i1}) \log &\left\{ \frac{\partial C[F_1(0; \vartheta_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}); \vartheta_i^{(c)}]}{\partial F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)})} \right\} \\ + y_{i1} \log &\left\{ 1 - \frac{\partial C[F_1(0; \vartheta_i^{(1)}), F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}); \vartheta_i^{(c)}]}{\partial F_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)})} \right\} + \log \left\{ f_2(y_{i2}; \boldsymbol{\vartheta}_i^{(2)}) \right\} \end{aligned} \tag{5}$$

The link function for the binary margin can be set to logit, probit, or cloglog.

2.3 Estimation via component-wise gradient boosting

As mentioned earlier and further emphasized by the summation index $P_k^{(\bullet)}$ shown in equation (2), there may not be strong a priori evidence of which subset of covariates (or if any at all) affects the individual parameters $\vartheta_k^{(\bullet)}$ of the bivariate distribution $F(\cdot, \cdot; \boldsymbol{\vartheta})$. Therefore, we resort to component-wise gradient boosting or statistical boosting[25,26] to estimate all coefficients simultaneously. While boosting is a general concept from machine learning, it has also been extended towards estimating statistical models.[27] The term *component-wise* highlights that this particular boosting framework fits the base-learners (components) one-by-one and greedily updates the model by updating only the best-performing component.[28] In our case, the base-learners are the additive components $s_{rk}^{(\bullet)}(x)$ in equation (2). We refer to Hothorn et al.[28] and Mayr et al.[29] for a complete list of the currently implemented base-learners. Estimating the model coefficients corresponds to solving the optimization problem:

$$\hat{\boldsymbol{\eta}} = \arg\min_{\eta} \Big[E_{\boldsymbol{Y}} \left\{ \omega \left(\boldsymbol{Y}; \boldsymbol{\eta} \right) \right\} \Big]$$

where the vector $\boldsymbol{\eta} = (\boldsymbol{\eta}^{(1)}, \boldsymbol{\eta}^{(2)}, \boldsymbol{\eta}^{(c)})^\top \in \mathbb{R}^K$ contains all additive predictors corresponding to the parameters of the bivariate distribution and $\hat{\boldsymbol{\eta}}$ denotes their estimates. The term $\omega(\cdot)$ represents the loss function, which in our case corresponds to the negative log-likelihood of the regression model, that is, $\omega(\cdot) = -\ell(\cdot)$. In general, minimizing the expectation of the loss is intractable. In practice, given a sample of $i = 1, \ldots, n$ observations, one minimizes the *empirical risk* $\rho = (1/n) \sum_{i=1}^n \omega(\boldsymbol{y}_i; \boldsymbol{\eta}_i)$ iteratively. In each boosting iteration, the algorithm fits each of the pre-specified base-learners in each predictor individually to the negative gradient of the loss function (also sometimes referred to as *pseudo-residuals)*, that is, $-\partial\rho/\partial\eta_k^{(\bullet)}$. Only the best-fitting base-learner is selected and a 'weak' update of the model is conducted. The fitting procedure is run for a pre-specified number of iterations denoted by $\mathtt{m}_{\mathtt{stop}}$, which plays a similar role like the penalty parameter 'λ' of the least absolute shrinkage and selection operator (LASSO),[30] and acts as the main tuning parameter. In our case, we conduct non-cyclical updates,[31] which means that only one out of all additive predictors is updated per fitting iteration. Only the update which leads to the highest decrease in the empirical risk is updated. By conducting *early stopping*, that is, using $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}} < \mathtt{m}_{\mathtt{stop}}$ fitting iterations, some base-learners will effectively be left out of the model, since they were not selected in any iteration. Hence early stopping results in intrinsic, data-driven variable selection as well as shrinkage of covariate effects. Algorithm 1 in Supplemental material A details the procedure including a mechanism for tuning of $\mathtt{m}_{\mathtt{stop}}$.

The data-driven variable selection and regularization of effect estimates resulting from boosting with early stopping are particularly suitable for exploratory data analyses or prediction modelling. In such cases, boosting can provide valuable insights by automatically selecting relevant variables without requiring prior knowledge of their importance. However, it is worth noting that as a main limitation, statistical boosting in our flexible model class lacks the availability of asymptotic theory to, for example, construct confidence intervals or to conduct inference.

3 Simulation study

In this section, we summarize the main findings of our simulation study. We refer to Supplemental material B for all details of the simulation study. We consider three response scenarios in Sections SB1 to SB3 in Supplemental material B., one for bivariate binary, count and mixed outcomes each under different levels of sparsity. The main goals are to evaluate (i) estimation, (ii) variable selection and (iii) predictive performance of our proposed bivariate copula approach compared to the benchmark of estimating two separate (and thus independent) univariate models. Additionally, we investigate the performance of the out-of-sample negative log-likelihood evaluated on an additional test data set to identify the correct copula function in Section SB4 in Supplemental material B. The code used to reproduce the simulations can be found in the following repository: https://github.com/GuilleBriseno/BoostDistCopReg_BinDiscMix.

3.1 General settings

All boosting models are fitted using the `gamboostLSS` package. A training data set of $n_{\text{train}} = 1000$ observations and a fixed step-length of $\texttt{s}_{\texttt{step}} = 0.1$ for all distribution parameters are used. The stopping iteration $\texttt{m}_{\texttt{stop}}$ is optimized by minimising the out-of-bag negative log-likelihood using a validation data set with $n_{\texttt{mstop}} = 1500$ observations from the same underlying distribution (see Step 4 in Algorithm 1 in Supplemental material A). We apply L_2-stabilisation to the parameter-specific gradients in order to obtain similar step-lengths among the various dimensions of the model, see Hofner et al.[20] for details on gradient stabilisation. The performance of the copula and univariate models is evaluated using multivariate proper scoring rules (negative log-likelihood and energy score[32]), both oriented such that lower values indicate better performance and evaluated on an additional test data set of size $n_{\text{test}} = 1000$ observations that are not used in the fitting process or for tuning. The energy score is computed using the `scoringRules`[33] package. We include univariate distribution-specific evaluation criteria as well, although we remark that these criteria do not take the dependence between the responses into account. For binary responses, we use the Brier score and the area under the curve. For the remaining discrete and mixed responses we compute the univariate mean squared error of prediction comparing the true Y_j with its prediction $\hat{Y}_j, j = 1, 2$. The bivariate observations are generated using the `VineCopula`[34] package. All performance measures are averages over the observations in the test set and averages over 200 independent data set replications. Lastly, we report the selection rates of the informative and non-informative variables for each distributional parameter. The selection rates are defined as the percentage of simulation replications in which the informative/non-informative variables have been selected, averaged by the number of informative/non-informative variables in each distribution parameter, respectively.

3.2 Details of sparsity and dimensions

To challenge the boosting algorithm, we consider different amounts of sparsity and covariates that are informative in more than one distribution parameter. For the bivariate binary response scenario (Section SB1 in Supplemental material B) $p_1 = 100, p_2 = 100$, and $p_3 = 1000$ candidate covariates are considered. Only six covariates have a linear effect on the bivariate distribution, whereas the rest are noise variables. This leads to 50% (p_1), 5% (p_2) and 0.5% (p_3) of the candidate covariates being informative, respectively, thereby reflecting low, medium and high levels of sparsity. The scenario with bivariate counts (Section SB2 in Supplemental material B) is comprised of linear and non-linear data-generating processes (DGPs) with $p_1 = 10$ independent variables. In these configurations 60% (linear DGP) and 50% (non-linear DGP) of the covariates were informative. The mixed binary & continuous scenario (Section SB3 in Supplemental material B) consists of linear and a non-linear DGPs with $p_1 = 10$. In those simulations, 50% (linear DGP) and 30% (non-linear DGP) of the covariates were informative. With $n_{\text{train}} = 1000$ throughout, all but the SB1/p_3 case, where $p = n$, are low-dimensional settings with $p < n$.

3.3 Overall summary of simulation results

In general, the performance of the proposed boosted copula models is satisfactory. They effectively detect and recover all effects across different parameters of the bivariate distribution. Notably, the copula dependence parameter shows a stronger shrinkage of informative effects compared to other parameters. As the number of considered covariates increases,

the degree of shrinkage also rises. This behaviour may be attributed to the greedy nature of the algorithm, since a reduction of the loss from including a covariate with a small coefficient in the dependence parameter might not be large enough compared to updating a coefficient in any other parameter corresponding to the margins. Consequently, this can lead to sparser dependence parameters with relatively small effects being falsely disregarded. The choice of $\texttt{m}_{\texttt{stop}}$ in the distribution parameter of the copula remains an under-explored area, deserving attention in future research to address this issue.

Overall, the copula approach is competitive in terms of selection rates of covariates in the marginal parameters and satisfactory in identifying the most relevant effects in the dependence parameter. Based on scores evaluating the predictive behaviour of the joint distribution, the added value compared to using boosting with independent univariate models becomes obvious even for moderate associations between the response components.

4 Biomedical applications

In this section, we illustrate the versatility of our proposed boosted distributional copula regression approach by analysing three different biomedical research questions. In Section 4.1, we model the joint distribution of two binary responses which correspond to the presence of heart disease (yes/no) as well as the presence of high cholesterol (yes/no) using data from the large-scale biomedical database UK Biobank genetic cohort study[35] under application number 81202. This corresponds to a high-dimensional setting in the covariate space. In Section 4.2, we are concerned with the joint distribution of a bivariate count vector comprised of the number of doctor consultations and the number of prescribed medications from Australian healthcare recipients using data from the R package `bivpois`.[36] We demonstrate how to conduct model-building when the choice of marginal distributions, as well as copula functions, is not clear. Lastly, in Section 4.3, we investigate the distribution of two mixed responses relevant for analysing infant undernutrition in India emanating using data from the Demographic and Health Survey (DHS; https://dhsprogram.com, accessed on 13 December 2023).[37] In what follows, the step-length of the boosting algorithm is set to $\texttt{s}_{\texttt{step}} = 0.1$ and the number of fitting iterations $\texttt{m}_{\texttt{stop}}$ is optimized via the predictive or out-of-bag risk as outlined in Step 4 of Algorithm 1 in Supplemental material A. We resort to L_2-stabilisation in order to achieve similar effective step-lengths across the different parameters of the bivariate distributions.

4.1 Chronic ischaemic heart disease and high cholesterol

We analyse a subsample consisting of $n = 30{,}000$ individuals and $p = 1867$ pre-filtered genetic variants (covariates). This sample has been previously analysed in Strömer et al.[17] using a bivariate Bernoulli distribution. The responses are the presence of chronic ischaemic heart disease (`CIHD`), and high cholesterol (`cholesterol`), both encoded as binary variables. The prevalence of the two factors in our sample is 7.2% and 32.3%, respectively.

4.1.1 Model specification

We build the joint distribution using a Gaussian copula with logit margins. We split the sample into two partitions dedicated for fitting ($n_{\text{train}} = 20{,}000$) and tuning of $\texttt{m}_{\texttt{stop}}$ ($n_{\texttt{mstop}} = 10{,}000$). The additive predictors of the bivariate distribution are

$$\eta_{i1}^{(\bullet)} = \beta_{01}^{(\bullet)} + \sum_{r=1}^{1{,}867} \beta_{r1}^{(\bullet)} x_{ir}, \quad \text{with } \bullet = \{1, 2, c\}$$

4.1.2 Results

The estimated coefficients, expressed as the exponential absolute values in each margin and the dependence parameter, are shown in a Manhattan-type plot[38] in Figure 2. The scale of the y-axis of the Manhattan plot has been modified to reflect the importance of the different genetic variants via the exponential absolute value of the estimates coefficients (for the margins, similar to an odds ratio from logistic regression). Using the estimated dependence parameters $\hat{\vartheta}_i^{(c)}$, for $i = 1, \ldots, n$, we compute the corresponding Kendall's τ, which range from $\hat{\tau} \in [-0.567;\ 0.289]$. This result indicates that there is a moderate negative dependence between the probabilities of chronic heart disease and high cholesterol. This finding most likely reflects the common use of statins in the population of patients already diagnosed with chronic heart disease.[39] Our proposed boosting method selects several variants in the respective parameters of the bivariate distribution. For instance, out of a potential 1867 possible candidates, 140 variants are selected in the first margin ($\vartheta_1^{(1)}$), 322 variants in the second margin ($\vartheta_1^{(2)}$) and 181 in the dependence parameter $\vartheta^{(c)}$ with some overlap in the selected variants between the parameters (90 variants selected for two out of three parameters). A total of 19 variants are shared between the dependence parameter and $\vartheta_1^{(1)}$, whereas $\vartheta_1^{(2)}$ and $\vartheta^{(c)}$ have 48 variants in common. Moreover, 23 variants are shared among the margins. The findings of our copula model agree with previous studies on the location of cholesterol-associated genes, see, for example, Richardson et al.,[40] where the highest estimated coefficient values are present.

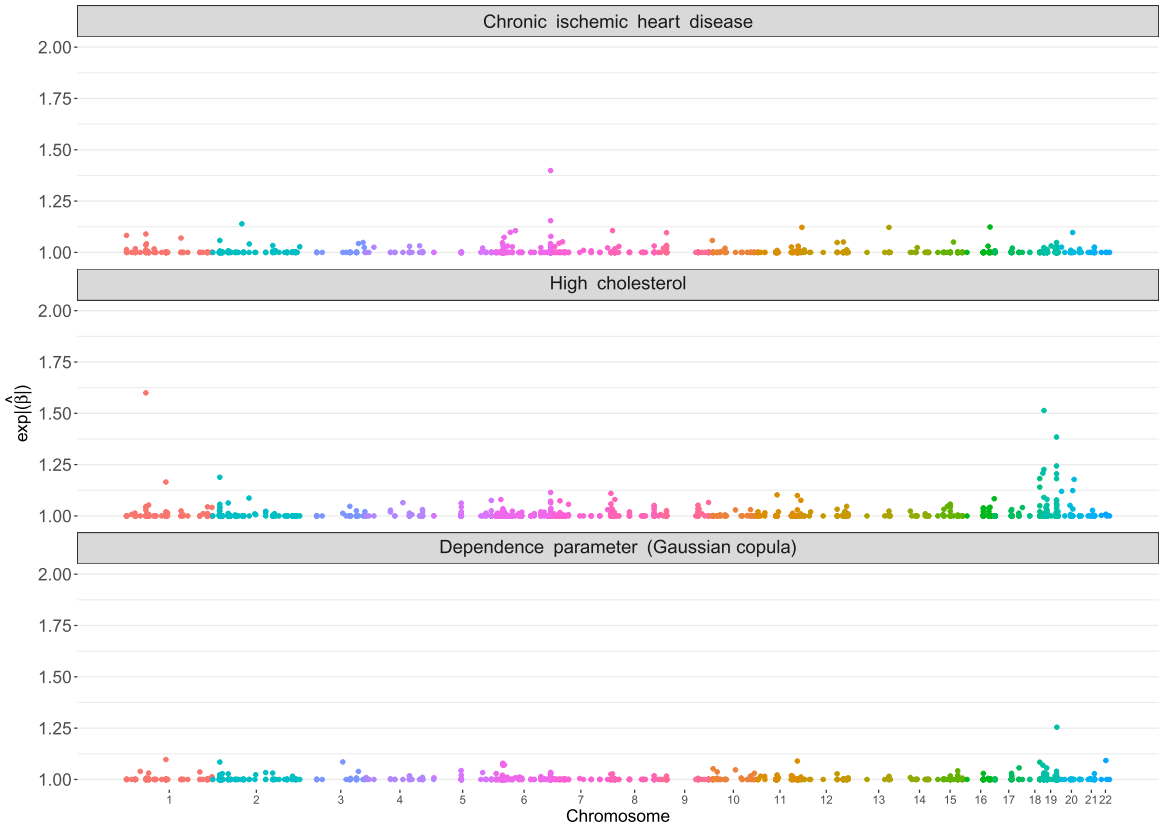

Figure 2. Application in Section 4.1. Manhattan-type plots of the estimated coefficients (expressed in exponential absolute values of the estimated values) of the boosted bivariate binary model using a Gaussian copula. The *x*-axis represents the genomic location of the variants and the *y*-axis shows $\exp(|\hat{\beta}_j|)$, $j = 1, \ldots, p$.

4.2 Doctor consultations and prescribed medications in Australia

We study the joint distribution of a bivariate count response comprised of the number of doctor consultations (`doctorco`) and the number of prescribed medications (`prescrib`) of healthcare recipients from Australia. The sample consists of $n = 5190$ observations and we use 65% of them to fit the model ($n_{\text{train}} = 3114$), and 25% for optimising $\texttt{m}_{\texttt{stop}}$ ($n_{\text{mstop}} = 1298$). An additional test partition of $n_{\text{test}} = 778$ observations is used to determine the best-fitting marginal distributions and copula function. The dataset comprises two continuous covariates. These are `age` (age in years divided by 100) and `income` (annual income in Australian dollars divided by 1000). In addition, the binary covariate `gender` (1 female, 0 male) is reported.

4.2.1 Marginal distributions

The best-fitting marginal distributions have been determined via the out-of-sample negative log-likelihood on the test partition of the data, see Table SC1 in Supplemental material C for more details. As shown in Figure 1(b), each of the marginal responses exhibits a large amount of zeros and their respective variances differ from the mean ($\overline{\texttt{doctorco}} = 0.302$; $Var(\texttt{doctorco}) = 0.637$, and $\overline{\texttt{prescrib}} = 0.863$; $\text{Var}(\texttt{prescrib}) = 2.003$). While these descriptive statistics do not account for the covariates, we also find that with regressors the Poisson distribution is not suited to model the conditional distribution of the two responses. The best-fitting marginal distributions in terms of the out-of-sample negative log-likelihood are the zero-altered logarithmic distribution $(\vartheta_1^{(1)}, \vartheta_2^{(1)})^\top$ for `doctorco`, where the probability of observing a zero is modelled by the parameter $\vartheta_2^{(1)}$. The zero-inflated negative binomial distribution $(\vartheta_1^{(2)}, \vartheta_2^{(2)}, \vartheta_3^{(2)})^\top$ is the most suitable for `prescrib`. With this, the probability of observing a zero is explicitly modelled via the parameter $\vartheta_3^{(2)}$.

4.2.2 Copula selection

The copula was selected by means of the out-of-sample negative log-likelihood (using the same test data as for the margins) out of six possible candidates: Gaussian, Frank, Clayton, Gumbel, Farlie–Gumbel–Morgenstern and Ali–Mikhail–Haq

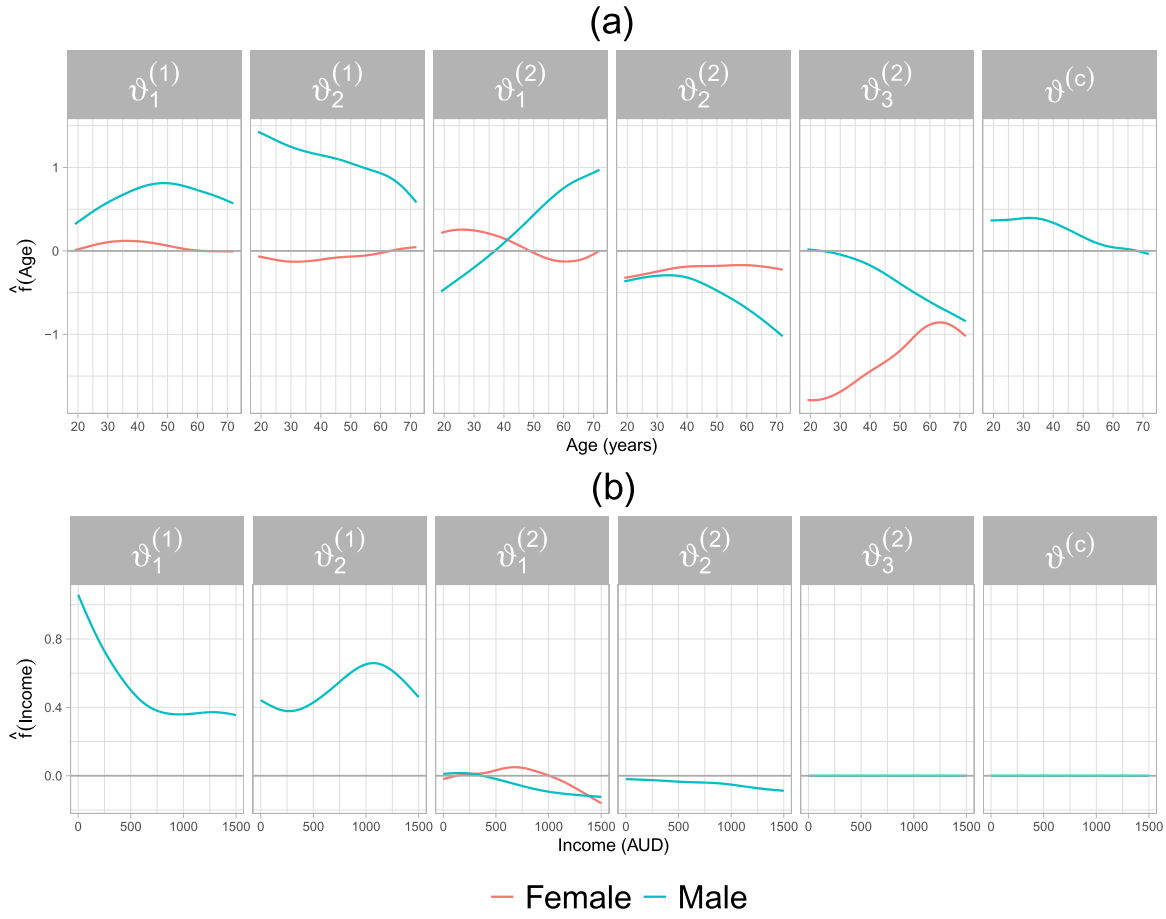

Figure 3. Application in Section 4.2. Estimated partial effects of age (a) and income (b) on the additive predictors $\eta_k^{(\bullet)}$ of the parameters of the marginal distributions as well as the dependence parameter of a Clayton copula.

copulas, with the Clayton copula giving the best out-of-sample negative log-likelihood. This indicates that the data support the presence of lower tail dependence, that is, strong dependence of very low values in both marginal responses. In addition, the Clayton copula performs better than independent margins as well as the bivariate Poisson distribution, see Table SC2 in Supplemental material C for more details.

4.2.3 Predictor specification

As a result of the selection of marginal distributions, there are six parameters in the bivariate distribution ($K_1 = 2$, $K_2 = 3$, $K_c = 1$) and all additive predictors in the distribution share the following configuration:

$$\begin{aligned}\eta_{ik}^{(\bullet)} = \beta_{0k}^{(\bullet)} + \beta_{1k}^{(\bullet)}\texttt{gender}_i + s_{1k}^{(\bullet)}\left(\texttt{income}_i\right) : \texttt{gender}_i \\ + \quad s_{2k}^{(\bullet)}\left(\texttt{age}_i\right) : \texttt{gender}_i + s_{3k}^{(\bullet)}\left(\texttt{income}_i, \texttt{age}_i\right), \\ \forall k = 1, \ldots, K_\bullet, \ \bullet \in \{1, 2, c\}\end{aligned}$$

where the term $s(\cdot) : \texttt{gender}$ denotes a varying coefficient term, where `age` or `income` are effect modifiers of the effect of `gender`, respectively. The base-learner $s_3^{(\bullet)}(\texttt{income}, \texttt{age})$ indicates the interaction term of the respective covariates and is specified as a two-dimensional P-spline. Therefore, this configuration takes into account the main effect of all covariates in the data as well as possible interactions between them.

4.2.4 Results

Table SC3 in Supplemental material C shows the base-learners selected in each parameter of the joint bivariate distribution. The fitted values of the dependence expressed as Kendall's τ range within $\hat{\tau} \in [0.341;\ 0.539]$, indicating a moderate

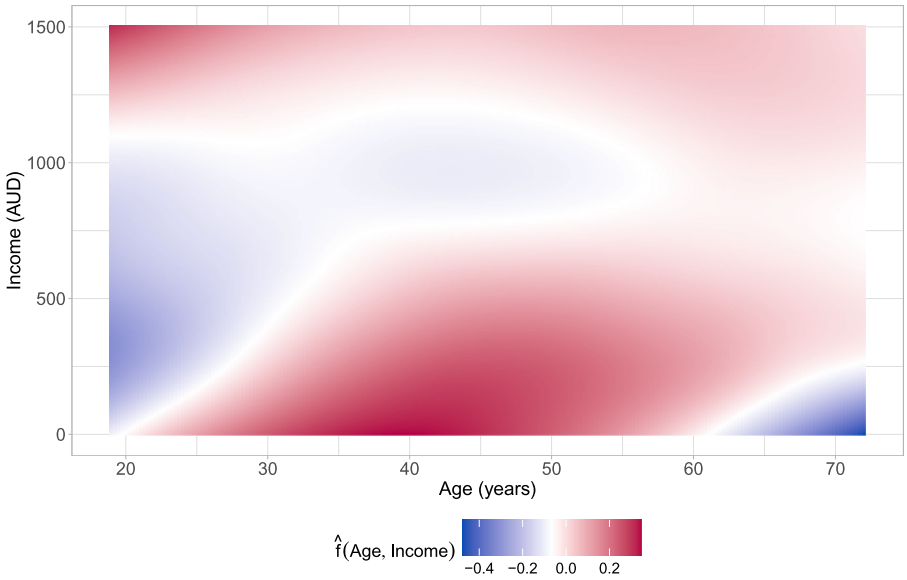

Figure 4. Application in Section 4.2. Estimated partial interaction effect of age and income on the additive predictor $\eta_1^{(2)}$ of the parameter $\vartheta_1^{(2)}$ (number of prescribed medications; `prescrib`).

estimated dependence between the margins in the sample, conditional on all selected covariate effects. Only the main effect of `age` and `gender` were selected on the copula dependence parameter.

The results of non-linear effect estimates and selected effect modifiers are depicted in Figure 3. The covariate `age` has a non-zero effect in all parameters of the bivariate distribution (see panel (a) of Figure 3) and it interacts with `gender` only on the marginal distributions. In particular, the effect of age on $\vartheta_1^{(1)}$ is increasing between 20 and 50 years, and then becomes decreasing for older male individuals. For female individuals, the effect follows a similarly shaped pattern, albeit the positive effect lasts until the mid-30s and the range of the effect is close to zero. Increasing age leads to smaller values of the parameter $\vartheta_2^{(1)}$ for male individuals, whereas for females the effect leads to an increase in $\eta_2^{(1)}$ but its range is once again close to zero. The two aforementioned parameters jointly determine the expectation and variance of `doctorco`, whereas the parameter $\vartheta_2^{(1)}$ explicitly models the probability of observing a zero. Hence, for older individuals, it becomes less likely to have zero doctor consultations.

For male individuals, increasing age leads to an increase in the predictor of $\vartheta_1^{(2)}$, which partially determines the expected number of prescribed medications. Intuitively, the predictor of $\vartheta_3^{(2)}$ decreases almost linearly with the male individual's age, which directly translates to the logit of a decreased probability of observing a zero in the second margin. In other words, older male individuals are more likely to have a number of prescribed medications that are larger than zero. Conversely, the effect of age of female individuals on the predictor of $\vartheta_3^{(2)}$ shows an upward trend, which indicates an increasing likelihood of having zero prescribed medications. A downward-sloping effect of age is estimated for the parameter $\vartheta_2^{(2)}$ for both female and male individuals. Additionally, the dependence between the margins decreases in older individuals as seen in the panel corresponding to $\vartheta^{(c)}$. Note that the interaction between age and gender is not selected in the dependence parameter.

The covariate `income` is selected in four parameters of the bivariate distribution, see Figure 3(b). The individual's income has a non-zero effect on the parameters of `doctorco` distribution and shows no interaction with `gender`. Conversely, `income` exhibits a much smaller, albeit downward-sloping, effect on the parameters $\vartheta_1^{(2)}$ and $\vartheta_2^{(2)}$ of the distribution of `prescrib`. The interaction of `income` and `gender` is selected only on the parameter $\vartheta_1^{(2)}$. The covariate `income` was neither selected on the parameter $\vartheta_3^{(2)}$ nor on the dependence parameter. This result indicates that `income` does not play a role in the association between `prescrib` and `doctorco`. The interaction between `age` and `income` is only selected for the parameter $\vartheta_1^{(2)}$. The estimated two-dimensional P-spline depicted in Figure 4 shows that there is an interplay between an individual's `age` and `income` on the expected number of prescribed medications (`prescrib`). For younger individuals with low to moderate income, the interaction reduces the value of the additive predictor of $\vartheta_1^{(2)}$. A similar pattern can be observed for individuals in a higher age bracket ($\geq$ 70 years) with low income.

The covariate `gender` was selected in all parameters except for $\vartheta_1^{(1)}$, see Table 1, middle block. The estimates of `gender` in the first margin indicate that the expected value of both responses is higher for female healthcare recipients, *ceteris paribus*. This is due to $\vartheta_2^{(1)}$ directly modelling the probability of observing no doctor visits. The estimated effect of `gender`

Table 1. Estimated linear effects for applications in Sections 4.1) (first block), 4.2 (second block) and 4.3 (third block) across distribution parameters.

		Margin 1		Margin 2			Copula
Application	Covariate	$\vartheta_1^{(1)}$	$\vartheta_2^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta_3^{(2)}$	$\vartheta^{(c)}$
Bivariate binary		Bernoulli (logit)		Bernoulli (logit)			Gaussian
	`Intercept`	−1.198	–	−0.317	–	–	0.442
Bivariate count		ZALG		ZINBI			Clayton
	`Intercept`	−1.193	−0.050	−0.255	0.234	0.022	0.452
	`gender` (female)	0	−0.218	0.189	−0.447	−1.171	−0.379
Bivariate mixed		Bernoulli (probit)		Gaussian			Clayton 270°
	`Intercept`	−0.230	–	0.003	0.008	–	0
	`cgender` (female)	−0.031	–	0	0.002	–	0

The symbol "–" indicates that the distribution does not feature the respective parameter, whereas 0 indicates that the algorithm did not select the respective covariate.

in $\vartheta_3^{(2)}$ also suggests that the probability of having zero prescribed medications is lower for female recipients compared to male individuals. Lastly, the dependence between the margins is lower for female individuals, relative to their male counterparts.

4.3 Determinants of infant undernutrition in India

We analyse a sample of $n = 24{,}286$ observations to study jointly two determinants of child undernutrition in India. The binary response `fever` indicates whether a child has had fever up to two weeks prior to the survey interview, whereas `wasting` denotes low weight-for-height, indicating an acute recent weight loss. According to UNICEF, this is the most immediate, visible and life-threatening form of undernutrition.[41] The individuals in the sample are spread across 438 administrative units (districts) with some imbalance in the number of observations per district. We resort to a slightly different sub-sampling scheme compared to the previous applications in order to obtain n_{train}, and $n_{\texttt{mstop}}$. We include all observations corresponding to districts with a sample size below or equal to 40 in n_{train}. For all other districts with more than 40 observations, we sample without replacement and obtain a fraction of around 75% of the total observations used for training ($n_{\text{train}} = 18{,}214$) and 25% for optimizing `mstop` ($n_{\texttt{mstop}} = 6072$). Table SC4 (Supplemental material C) summarizes the responses and available covariates.

4.3.1 Model specification

We follow Klein et al.[15] and set the link function for the model of `fever` to probit, whereas for `wasting` we resort to a heteroscedastic Gaussian distribution. The dependence between the margins is modelled using a Clayton copula rotated by 270°. This allows us to model dependence between very high values of `fever` and very low values of `wasting`. It seems reasonable to expect such a dependence structure to be supported by the data, since it is likely that the probability of children experiencing fever is prone to be dependent on low weight-for-height values (`wasting`, i.e. undernourished infants). Consequently, the bivariate distribution has $K = 4$ distribution parameters. In Klein et al.,[15] the additive predictor of the margins was fixed and an information-criterion-based model selection procedure was conducted using different configurations of the predictor of $\vartheta^{(c)}$. Here we allow our proposed approach to select the variables in all predictors of the bivariate distribution in a data-driven manner, without further input from the analyst. That is,

$$\begin{aligned}\eta_{ik}^{(\bullet)} = \beta_{0k}^{(\bullet)} + \beta_{1k}^{(\bullet)}\texttt{cgender}_i + s_{1k}^{(\bullet)}\left(\texttt{cage}_i\right) + s_{2k}^{(\bullet)}\left(\texttt{mbmi}_i\right) \\ + \quad s_{3k}^{(\bullet)}\left(\texttt{breastfeeding}_i\right) + s_{4k}^{(\bullet)}\left(\texttt{distH}_i\right)\end{aligned}$$

where $k = 1, \ldots, K$, $\bullet \in \{1, 2, c\}$ and $s_{4k}^{(\bullet)}(\texttt{distH}_i)$ is set as a Markov random field base-learner to model the discrete spatial information of the districts in the data. The covariates `cage`, `mbmi` and `breastfeeding` are incorporated using P-spline base-learners with 20 knots and second-order difference penalties, whereas a linear base-learner is used for `cgender`.

4.3.2 Results

The estimated dependence between the margins in terms of Kendall's τ ranges within $\hat{\tau} \in [-0.561;\ -0.052]$, suggesting a negative dependence between `wasting` and `fever`. This is a reasonable finding since a lower `wasting` score implies a

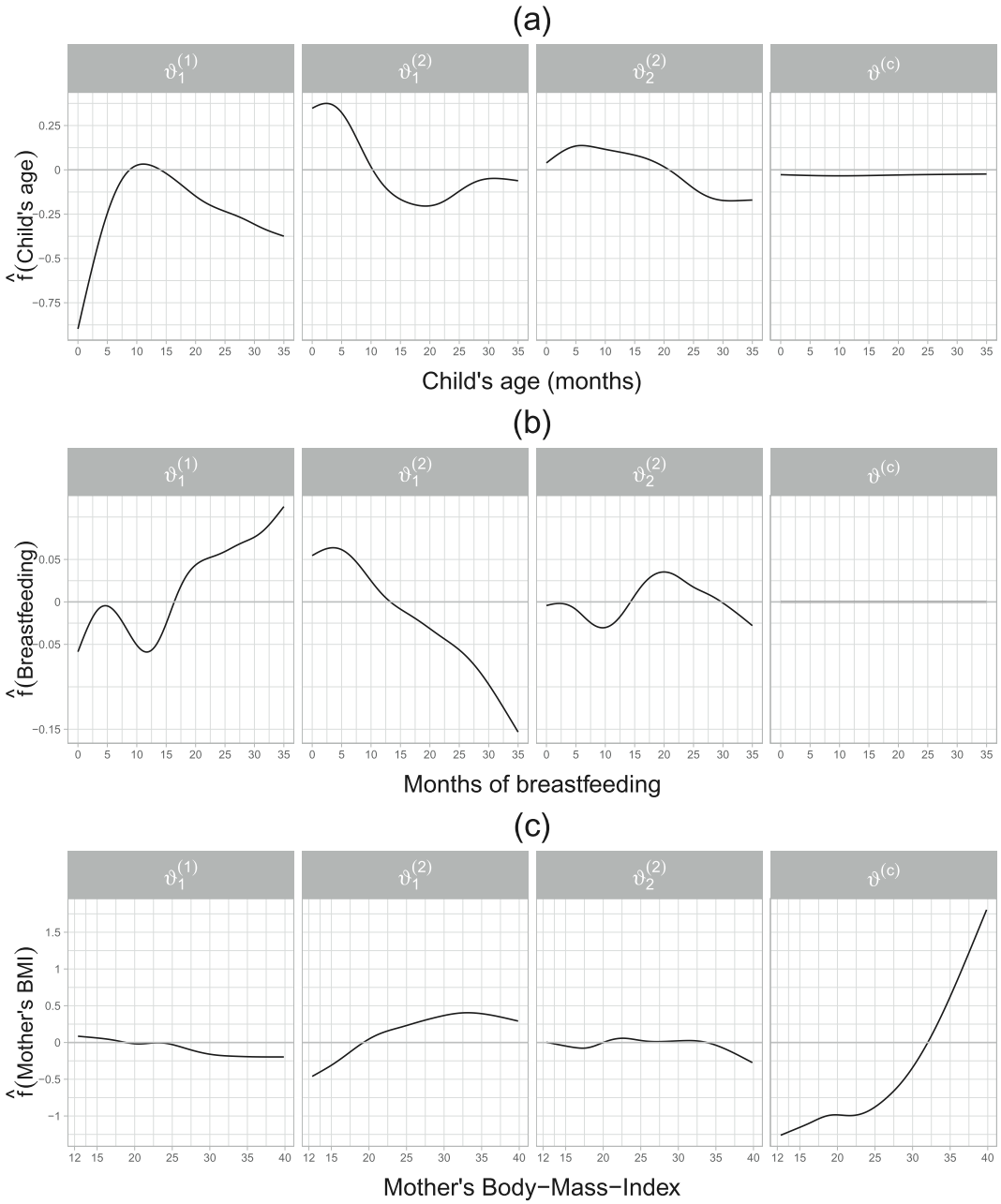

Figure 5. Application in Section 4.3. Estimated partial effects of the child's age (`cage`, a), months of breastfeeding (`breastfeeding`, b) and the mother's body-mass-index (`mbmi`, c) on the additive predictors $\eta_k^{(\bullet)}$ of the parameters of the margins as well as the dependence parameter of a Clayton copula rotated by 270°.

more severe form of undernutrition, whereas the risk of fever is expected to be positively associated with poor health status. The estimated non-linear effects of the covariates `cage`, `breastfeeding` and `mbmi` are visualized in Figure 5. It can be seen that children within 0 and $\approx$ 12 months of age have an increasing likelihood of `fever`. The estimated effect of `cage` is downward-sloping in the first 20 months on the expectation of `wasting`, whereas on the standard deviation, a similar pattern is observed albeit with a much smaller slope, see Figure 5(a). In terms of the dependence structure, the child's age appears to have a negligible effect. The estimated effect of `breastfeeding` on `fever` depicted in Figure 5(b) shows an upward slope and on $\vartheta_1^{(1)}$ a downward slope. The presence of breastfeeding at a later age of the child could reflect a lack of other sources of nourishment apart from the mother, serving as a proxy for household's poverty, thus driving the probability of `fever` upwards and the expected value of `wasting` downwards. The variable `breastfeeding` is not selected in the dependence

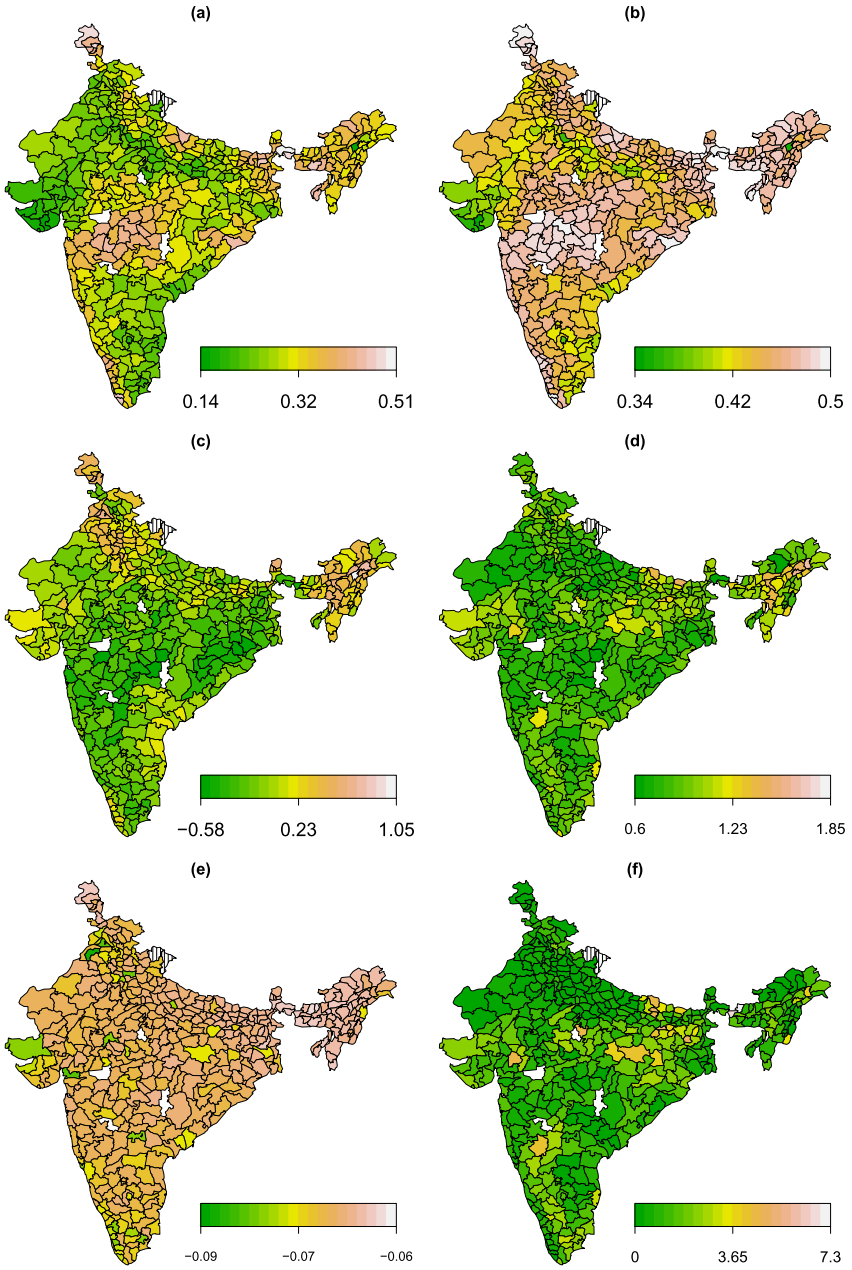

Figure 6. Application in Section 4.3. Shown are (a) expected value, (b) standard deviation of `fever`; (c) estimated expected value, (d) standard deviation of `wasting`; (e) estimated Kendall's τ, and (f) joint probabilities in % of having fever and moderate undernutrition according to the Clayton copula rotated by 270°.

parameter. Compared to `cage` and `breastfeeding`, the mother's body-mass-index (`mbmi`) shows a small to moderate (see $\vartheta_1^{(2)}$) association with the margins, see Figure 5(c). The effect of `mbmi` is slightly increasing in the expectation of `wasting` and remains stable at around `mbmi` ≈ 25. However, the effect of `mbmi` leads to a sharp increase in the dependence between the margins after it reaches values of approximately 25. The covariate `cgender` was not selected in $\vartheta_1^{(2)}$ as well as $\vartheta^{(c)}$ and it shows a very small value in $\vartheta_2^{(2)}$, compare Table 1, third block. Finally, Figure 6 presents various estimated quantities (expectation, standard deviation and Kendall's τ, joint probabilities) according to the spatial structure of the data. The spatial component modelling the districts (`distH`) is selected in all parameters. In Figure 6(a) it can be observed that the districts located in the centre of India exhibit a higher probability of `fever`, however, the standard deviation of `fever` is rather high across the country (see Figure 6(b)). The expectation of `wasting` remains mostly low throughout all districts,

with some exceptions located in the north and north-eastern districts of India, see Figure 6(c). Compared to `fever`, the standard deviation of `wasting` is rather low in most districts, see Figure 6(d). Figure 6(e) and (f) visualizes the per district average of the estimated dependence between the margins in terms of Kendall's τ and the estimated joint probabilities (in %) of having fever and moderate undernutrition, that is, $P(Y_1 = 1, Y_2 < -2)$. It can be seen that the magnitude of the dependence is larger in some districts located in the north-western are, as well as the south-eastern coast of India. The joint probabilities of fever and moderate undernutrition indicate that children located in mid-eastern districts are more prone to suffer from undernutrition.

5 Discussion

We have extended the boosted distributional copula regression approach[16] to accommodate arbitrary response types on different domains. We conducted a wide range of simulation studies to investigate the predictive performance, as well as the estimation capabilities of our proposed method. Overall, we found that our approach outperforms univariate boosting models when it comes to probabilistic forecasting for the joint bivariate distribution.

We were able to demonstrate that our proposed copula approach allows us to capture the nuances of each marginal response, such as zero-inflation, over-dispersion, or heteroscedasticity, while also modelling the dependence between the margins using only one statistical model. Additionally, our methodology and software implementation allow us to conduct data-driven variable selection without further input from the analyst as well as transparent and reproducible research.

We have illustrated the application of our approach on three diverse biomedical datasets. In the first application, we identified relevant genetic variants associated with the dependence of high cholesterol and ischaemic heart disease. Although not conducted here due to computation time constraints, other copula functions than the Gaussian copula could be tested in order to investigate whether the data support lower or upper tail dependence. In our second healthcare-related application, we found that data on the number of doctor consultations and number of prescribed medications support lower tail dependence, that is, dependence between extremely low values of the margins. Finally, in the third application, we studied the joint distribution of two determinants of infant undernutrition that emanate from different domains. One determinant is expressed as a binary indicator whereas the other is a continuous marker.

While our approach is very useful for conducting explanatory analyses and for predictive modelling, the main limitation of resorting to statistical boosting for model fitting is the lack of confidence intervals for the estimated effects. While in principle access to these is possible using bootstrap methods, doing so is a cumbersome and time-consuming task. Another limitation was observed in our simulation studies in Section 3. The boosted models have a tendency to select false positives throughout the fitting process and the different distribution parameters. Although the estimated effect of these false positives is in most cases small or negligible, a formal correction of these incorrectly estimated effects would be appealing. An adaptation of the de-selection procedure implemented by Strömer et al.[42] would lead to more sparse models and stable selection of informative covariates.

Another future field of application where data-driven variable selection can have a big impact is in observational studies where endogenous variables are present, see, for example, Briseño Sanchez et al.[43] and Wyszynski et al.[44] Statistical boosting could provide valuable insights in these scenarios, since the effect of endogenous variables is identifiable as long as so-called instruments are available, which boosting could help identify and to validate the analyst's beliefs. Lastly, we are also exploring an extension of our boosting methodology to fit distributional copula regression models for bivariate time-to-event data, which would greatly extend the applicability of our software implementation in biomedical research: In clinical applications, the interest may lie in overall survival expressed as a time of a landmark event (e.g. tumour progression), time of death, or another event time associated with a patient's condition or chronic disease. The issue in most time-to-event applications is the presence of censoring, posing a challenge for estimation. It could be argued that copula regression models for standard continuous outcomes would be applicable for time-to-event data, albeit under the absence of censoring – which is unrealistic in most practical scenarios. While some statistical packages such as `GJRM`[45] and `joint.Cox`[46] offer a wide range of functionality and flexibility for bivariate time-to-event data, but they lack the ability to conduct data-driven variable selection. Therefore, extending our proposed methodology also to multivariate censored time-to-event outcomes could help to fill this gap in medical research.

Acknowledgements

The authors gratefully acknowledge the computing time provided on the Linux HPC cluster at Technical University Dortmund (LiDO3), partially funded in the course of the Large-Scale Equipment Initiative by the DFG via project 271512359. The analysis of the UK Biobank was conducted under application number 81202.

Author Note

Hannah Klinkhammer and Andreas Mayr are now affiliated with the Department of Medical Biometry and Statistics, Philipps University of Marburg, Marburg, Germany.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: The work on this article was supported by the German research foundation (DFG) through the grants KL3037/2-1, MA7304/1-1 (428239776).

ORCID iDs

Guillermo Briseño Sanchez https://orcid.org/0000-0003-1303-7411
Nadja Klein https://orcid.org/0000-0002-5072-5347
Andreas Mayr https://orcid.org/0000-0001-7106-9732

Supplemental material

Supplemental material for this article is available online.

References

1. Klein N. Distributional regression for data analysis. *Annu Rev Stat Appl* 2024; **11**: 321–346.
2. Intemann T, Pohlabeln H, Ahrens DHW, et al. Estimating age- and height-specific percentile curves percentile curves for children using GAMLSS in the IDEFICS study. In: Wilhelm AF and Kestler HA (eds) *Analysis of large and complex data*. Cham: Springer International Publishing, 2016, pp.385–394.
3. Stasinopoulos MD, Rigby RA and Bastiani FD. GAMLSS: A distributional regression approach. *Stat Modell* 2018; **18**: 248–273.
4. Marra G and Radice R. Copula link-based additive models for right-censored event time data. *J Am Stat Assoc* 2020; **115**: 886–895.
5. Rigby RA and Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc Ser C: Appl Stat* 2005; **54**: 507–554.
6. Klein N, Kneib T, Klasen S, et al. Bayesian structured additive distributional regression for multivariate responses. *J R Stat Soc Ser C: Appl Stat* 2015; **64**: 569–591.
7. Craiu VR and Sabeti A. In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *J Multivar Anal* 2012; **110**: 106–120.
8. Yee TW. *Vector generalized linear and additive models*. New York: Springer, 2015.
9. Klein N and Kneib T. Simultaneous inference in structured additive conditional copula regression models: A unifying Bayesian approach. *Stat Comput* 2016; **26**: 841–860.
10. Nelsen RB. *An introduction to copulas*. New York: Springer, 2006.
11. Smith MS. Bayesian approaches to copula modelling. In: Damien P, Dellaportas P, Polson NG and Stephens DA (eds) *Bayesian theory and applications*. Oxford: Oxford University Press, 2013, pp.336–358.
12. Marra G and Radice R. Bivariate copula additive models for location, scale and shape. *Comput Stat Data Analy* 2017; **112**: 99–113.
13. Marra G and Radice R. A joint regression modeling framework for analyzing bivariate binary data in R. *Depend Model* 2017; **5**: 268–294.
14. van der Wurp H, Groll A, Kneib T, et al. Generalised joint regression for count data: A penalty extension for competitive settings. *Stat Comput* 2020; **30**: 1419–1432.
15. Klein N, Kneib T, Marra G, et al. Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Stat Med* 2019; **38**: 413–436.
16. Hans N, Klein N, Faschingbauer F, et al. Boosting distributional copula regression. *Biometrics* 2023; **79**: 2298–2310.
17. Strömer A, Klein N, Staerk C, et al. Boosting multivariate structured additive distributional regression models. *Stat Med* 2023; **42**: 1779–1801.
18. Norton EC, Dowd BE and Maciejewski ML. Odds ratios – current best practice and use. *JAMA* 2018; **320**: 84–85.
19. R Core Team. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing, 2022. https://www.R-project.org/.
20. Hofner B, Mayr A and Schmid M. gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *J Stat Softw* 2016; **74**: 1–31.
21. Nikoloulopoulos AK and Karlis D. Regression in a copula model for bivariate count data. *J Appl Stat* 2010; **37**: 1555–1568.
22. Marra G and Wyszynski K. Semi-parametric copula sample selection models for count responses. *Comput Stat Data Anal* 2016; **104**: 110–129.
23. Joe H. *Dependence modeling with copulas*. New York, NY: CRC Press, 2014.

24. Wyszynski K and Marra G. Sample selection models for count data in R. *Comput Stat* 2018; **33**: 1385–1412.
25. Friedman JH. Greedy function approximation: A gradient boosting machine. *Ann Stat* 2001; **29**: 1189–1232.
26. Bühlmann P and Hothorn T. Boosting algorithms: Regularization, prediction and model fitting. *Stat Sci* 2007; **22**: 477–505.
27. Mayr A, Binder H, Gefeller O, et al. The evolution of boosting algorithms: From machine learning to statistical modelling. *Methods Inf Med* 2014; **53**: 419–427.
28. Hothorn T, Bühlmann P, Kneib T, et al. Model-based boosting 2.0. *J Mach Learn Res* 2010; **11**: 2109–2113.
29. Mayr A, Fenske N, Hofner B, et al. Generalized additive models for location, scale and shape for high dimensional data: A flexible approach based on boosting. *J R Stat Soc Ser C: Appl Stat* 2012; **61**: 403–427.
30. Hepp T, Schmid M, Gefeller O, et al. Approaches to regularized regression: A comparison between gradient boosting and the LASSO. *Methods Inf Med* 2016; **55**: 422–430.
31. Thomas J, Mayr A, Bischl B, et al. Gradient boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Stat Comput* 2018; **28**: 673–687.
32. Gneiting T, Stanberry LI, Grimit EP, et al. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST* 2008; **17**: 211–235.
33. Jordan A, Krüger F and Lerch S. Evaluating probabilistic forecasts with scoringRules. *J Stat Softw* 2019; **90**: 1–37.
34. Nagler T, Schepsmeier U, Stoeber J, et al. VineCopula: Statistical inference of Vine Copulas, 2022. R package version 2.4.4. https://CRAN.R-project.org/package=VineCopula.
35. Bycroft C, Freeman C, Petkova D, et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* 2018; **562**: 203–209.
36. Karlis D and Ntzoufras I. Analysis of sports data by using bivariate Poisson models. *J R Stat Soc: Ser D (The Statistician)* 2003; **52**: 381–393.
37. Demographic and Health Survey, https://dhsprogram.com/Data/ (2023, accessed 13 December 2023).
38. Wang F. Chapter 3. Genome-wide association studies (GWAS): What are they, when to use them? In: Dluzen DF and Schmidt MH (eds) *Rigor and reproducibility in genetics and genomics. Translational and applied genomics*. San Diego, CA: Academic Press, 2024, pp.51–81.
39. Sinnott-Armstrong N, Tanigawa Y, Amar D, et al. Genetics of 35 blood and urine biomarkers in the UK Biobank. *Nat Genet* 2021; **53**: 185–194.
40. Richardson TG, Sanderson E, Palmer TM, et al. Evaluating the relationship between circulating lipoprotein lipids and apolipoproteins with risk of coronary heart disease: A multivariable Mendelian randomisation analysis. *PLoS Med* 2020; **17**: e1003062.
41. UNICEF. Nutrition and care for children with wasting, https://www.unicef.org/nutrition/childwasting#:text=Wasting%20is%20the%20most%20immediate,developmental%20delays%2C%20disease%20and%20death (2023, accessed 15 December 2023).
42. Strömer A, Staerk C, Klein N, et al. Deselection of base-learners for statistical boosting with an application to distributional regression. *Stat Methods Med Res* 2022; **31**: 207–224.
43. Briseño-Sanchez G, Hohberg M, Groll A, et al. Flexible instrumental variable distributional regression. *J R Stat Soc Ser A: Stat Soc* 2020; **183**: 1553–1574.
44. Wyszynski K and Marra G. Sample selection models for count data in R. *Comput Stat* 2017; **33**: 1385–1412.
45. Petti D, Eletti A, Marra G, et al. Copula link-based additive models for bivariate time-to-event outcomes with general censoring scheme. *Comput Stat Data Anal* 2022; **175**: 107550.
46. Emura T, Sofeu CL and Rondeau V. Conditional copula models for correlated survival endpoints: Individual patient data meta-analysis of randomized controlled trials. *Stat Methods Med Res* 2021; **30**: 2634–2650.

Supplementary Material

for

"Boosting distributional copula regression for bivariate binary, discrete and mixed responses"

Contents

Part A: Details on implemented boosting algorithm, copula functions, marginal distributions and negative gradients of the implemented loss functions.

Part B: Details and results of the simulation study.

Part C: Additional results and supporting information of the biomedical applications.

Part A

This supplement contains the detailed description of our boosting algorithm for distributional copula regression with faster tuning of the number of fitting iterations as well as the definition of implemented copulas and marginal distributions. In addition, it contains the tables that define the implemented copula functions and marginal distributions. Lastly, it shows the definition of the negative gradients of the implemented loss functions.

Algorithm 1 Non-cyclic boosting for distributional copula regression with faster tuning of fitting iterations $\mathtt{m_{stop}}$ by means of out-of-bag (*oobag*) risk.

Require:

Define the base-learners $b_r^{(\bullet)}(x_r)$ for $r = 1, \ldots, P_{vk}$, $\bullet = 1, 2, c$.
Set the step-length $\mathtt{s_{step}} \ll 1$ as well as the (non-optimal) number of fitting iterations $\mathtt{m_{stop}}$.
Set weights indicating the training and $\mathtt{m_{stop}}$-tuning partitions of the sample n_{train}, $n_{\mathtt{mstop}}$.
Set type of stabilisation to be applied to the negative gradient vector (L_2, median absolute deviation or none).
(1) Initialise all predictors $\hat{\eta}_k^{(\bullet)}$ corresponding to $\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}$ with offset values $\hat{\eta}_{k,[0]}^{(\bullet)}$.
for $m = 1, \ldots, \mathtt{m_{stop}}$ **do**
for $k = 1, \ldots, K$ in $\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}$ **do**
(a) Evaluate the parameter-specific negative gradient vector $-\boldsymbol{g}_{k,[m]}^{(\bullet)}$

$$-\boldsymbol{g}_{k,[m]}^{(\bullet)} = \left(-\boldsymbol{g}_{k,[m]}^{(\bullet)}(\boldsymbol{x}_i)\right)_{i=1,\ldots,n_{\text{train}}} = -\left(\frac{\partial \omega\left(\boldsymbol{y}_i, \hat{\boldsymbol{\eta}}_i\right)}{\partial \eta_k^{(\bullet)}}\bigg|_{\hat{\boldsymbol{\eta}}=\hat{\boldsymbol{\eta}}_{[m-1]}(\boldsymbol{x}_i)}\right)_{i=1,\ldots,n_{\text{train}}}.$$

(b) Fit $-\boldsymbol{g}_{k,[m]}^{(\bullet)}$ to each parameter-specific base-learner $b_{k,j}^{(\bullet)}(x_j)$.
(c) Select the best-fitting base-learner $\hat{b}_{k,j^\star}^{(\bullet)}$ via residual sum of squares criterion.

$$j^\star = \underset{j \in 1,\ldots P_k^{(\bullet)}}{\arg\min} \sum_{i=1}^{n_{\text{train}}} \left(-g_{k,[m]}^{(\bullet)}(\mathbf{x}_i) - \hat{b}_{k,j}^{(\bullet)}(x_i)\right)^2.$$

(d) Compute loss reduction of a weak update using $\hat{b}_{k,j^\star}^{(\bullet)}$.

$$\Delta\omega_{\vartheta_k^{(\bullet)}} = \sum_{i=1}^{n_{\text{train}}} \omega\left(\boldsymbol{y}_i; \hat{\eta}_k + \mathtt{s_{step}} \hat{b}_{k,j^\star}^{(\bullet)}(x_{ij^\star})\right).$$

end for
(2) Update the parameter with highest loss reduction $\vartheta_k^{(\bullet)^\star} = \arg\min_{\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}} \left(\Delta\omega_{\vartheta_k^{(\bullet)}}\right)$:

$$\hat{\eta}_{k,[m]}^{(\bullet)*}(\boldsymbol{x}_i) = \hat{\eta}_{k,[m-1]}^{(\bullet)*}(\boldsymbol{x}_i) + \mathtt{s_{step}} \cdot \hat{b}_{k,j^\star}^{(\bullet)}(x_{ij^\star}).$$

(3) For the remaining parameters $\vartheta_k^{(\bullet)} \neq \vartheta_k^{(\bullet)^\star}$, set $\hat{\eta}_{k,[m]}^{(\bullet)}(\boldsymbol{x}_i) = \hat{\eta}_{k,[m-1]}^{(\bullet)}(\boldsymbol{x}_i)$.
(4) Compute the out-of-bag risk at iteration $[m]$:

$$\text{risk}_{\text{oobag},[m]} = \sum_{i=1}^{n_{\mathtt{mstop}}} \hat{\omega}\left(\boldsymbol{y}_i; \hat{\boldsymbol{\eta}}_{\boldsymbol{i}}|_{\hat{\eta}=\hat{\eta}_{[m]}(\boldsymbol{x}_i)}\right).$$

end for
(5) Determine $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}}$ by means of the out-of-bag risk:

$$\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}} = \underset{m \in 1,\ldots,\mathtt{m_{stop}}}{\arg\min}\ \text{risk}_{\text{oobag},[m]}.$$

Table A1: Details of implemented copulas. The functions $\Phi_1^{-1}(\cdot)$ and $\Phi_2(\cdot)$ denote the quantile function and CDF of the univariate and bivariate standard normal distributions, respectively. Rotated copulas by 90, 180 and 270 degrees are respectively defined as: $C_{90} = F_2 - C(1-F_1, F_2; \vartheta^{(c)})$, $C_{180} = F_1 + F_2 - 1 + C(1-F_1, 1-F_2; \vartheta^{(c)})$ and $C_{270} = F_1 - C(F_1, 1-F_2; \vartheta^{(c)})$. The term $D_1(\vartheta^{(c)}) = \int_0^{\vartheta^{(c)}} \frac{t}{\exp(t)-1} dt$ is the Debye function and Φ_2 denotes the CDF of the bivariate Gaussian distribution with correlation coefficient $\vartheta^{(c)}$. Finally, AMH stands for Ali-Mikhail-Haq and FGM stands for Farlie-Gumbel-Morgenstern.

Copula	$C(F_1, F_2; \vartheta^{(c)})$	Range of $\vartheta^{(c)}$	Link	Kendall's τ
Gauss	$\Phi_2(\Phi_1^{-1}(F_1), \Phi_1^{-1}(F_2); \vartheta^{(c)})$	$\vartheta^{(c)} \in [-1, 1]$	$\tanh^{-1}(\vartheta^{(c)})$	$\frac{2}{\pi}\arcsin(\vartheta^{(c)})$
Clayton	$(F_1^{-\vartheta^{(c)}} + F_2^{-\vartheta^{(c)}} - 1)^{-1/\vartheta^{(c)}}$	$\vartheta^{(c)} \in (0, \infty)$	$\log(\vartheta^{(c)})$	$\frac{\vartheta^{(c)}}{\vartheta^{(c)}+2}$
Gumbel	$\exp\left[-\left\{(-\log(F_1))^{\vartheta^{(c)}} + (-\log(F_2))^{\vartheta^{(c)}}\right\}^{\frac{1}{\vartheta^{(c)}}}\right]$	$\vartheta^{(c)} \in [1, \infty)$	$\log(\vartheta^{(c)} - 1)$	$1 - \frac{1}{\vartheta^{(c)}}$
Frank	$-\vartheta^{(c)^{-1}} \log\Big(1 + (\exp(-\vartheta^{(c)}F_1) - 1)\cdot$ $(\exp(-\vartheta^{(c)}F_2) - 1)/(\exp(-\vartheta^{(c)}) - 1)\Big)$	$\vartheta^{(c)} \in \mathbb{R} \setminus \{0\}$	$\vartheta^{(c)}$	$1 - \frac{4}{\vartheta^{(c)}}[1 - D_1(\vartheta^{(c)})]$
AMH	$F_1F_2/(1 - \vartheta^{(c)}(1-F_1)(1-F_2))$	$\vartheta^{(c)} \in [-1, 1]$	$\tanh^{-1}(\vartheta^{(c)})$	$1 - \frac{2}{3}\vartheta^{(c)^2}(\vartheta^{(c)} + (1-\vartheta^{(c)})^2\log(1-\vartheta^{(c)}))$
FGM	$F_1F_2/(1 + \vartheta^{(c)}(1-F_1)(1-F_2))$	$\vartheta^{(c)} \in [-1, 1]$	$\tanh^{-1}(\vartheta^{(c)})$	$\frac{2}{9}\vartheta^{(c)}$
Joe	$1 - ((1-F_1)^{\vartheta^{(c)}} + (1-F_2)^{\vartheta^{(c)}} - (1-F_1)^{\vartheta^{(c)}}(1-F_2)^{\vartheta^{(c)}})^{(1/\vartheta^{(c)})}$	$\vartheta^{(c)} \in [1, \infty)$	$\log(\vartheta^{(c)} - 1)$	$1 + \frac{4}{\vartheta^{(c)^2}}\int_0^1 x\log(x)(1-x)^{2(1-\vartheta^{(c)})/\vartheta^{(c)}} dx$

Table A2: Details of newly implemented univariate marginal distributions for binary, continuous and discrete responses to be used together with copulas in the `gamboostLSS` package. For the Zero-Altered Logarithmic distribution, the term $\alpha = -[\log(1-\vartheta_1)]^{-1}$. For the Zero-Altered Negative Binomial distribution, the term $c = \frac{1-\vartheta_3}{\left(1-(1+\vartheta_1\vartheta_2)^{-1/\vartheta_2}\right)}$. All distributions use the parameterisation from Rigby et al. (2019).

Distribution	$E(Y)$	$Var(Y)$	Parameters & range	Links	Response range
Bernoulli	ϑ_1	$\vartheta_1\,(1-\vartheta_1)$	$\vartheta_1 \in [0,1]$	logit probit cloglog	$y \in \{0,1\}$
Gaussian	ϑ_1	ϑ_2^2	$\vartheta_1 \in \mathbb{R}, \vartheta_2 > 0$	Identity, log	$y \in \mathbb{R}$
Poisson	ϑ_1	ϑ_1	$\vartheta_1 > 0$	log	$y \in \mathbb{N}_+$
Geometric	ϑ_1	$\vartheta_1 + \vartheta_1^2$	$\vartheta_1 > 0$	log	$y \in \mathbb{N}_+$
Negative Binomial (I)	ϑ_1	$\vartheta_1 + \vartheta_2\vartheta_1^2$	$\vartheta_2,\ \vartheta_2 > 0$	log, log	$y \in \mathbb{N}_+$
Zero-Altered Logarithmic	$\frac{(1-\vartheta_2)\alpha\vartheta_1}{(1-\vartheta_1)}$	$\frac{(1-\vartheta_2)\alpha\vartheta_1(1-(1-\vartheta_2)\alpha\vartheta_1)}{(1-\vartheta_1)^2}$	$\vartheta_1,\ \vartheta_2 \in [0,1]$	logit, logit	$y \in \mathbb{N}_+$
Zero-Inflated Poisson	$(1-\vartheta_2)\vartheta_1$	$\vartheta_1(1-\vartheta_2)(1+\vartheta_1\vartheta_2)$	$\vartheta_1,\ \vartheta_2 \in [0,1]$	log, logit	$y \in \mathbb{N}_+$
Zero-Altered Negative Binomial	$c\vartheta_1$	$c\vartheta_1 + c\vartheta_1^2(1+\vartheta_2-c)$	$\vartheta_1, \vartheta_2 > 0,\ \vartheta_3 \in [0,1]$	log, log, logit	$y \in \mathbb{N}_+$
Zero-Inflated Negative Binomial	$(1-\vartheta_3)\vartheta_1$	$\vartheta_1(1-\vartheta_3) + \vartheta_1^2(1-\vartheta_3)(\vartheta_2+\vartheta_3)$	$\vartheta_1, \vartheta_2 > 0,\ \vartheta_3 \in [0,1]$	log, log, logit	$y \in \mathbb{N}_+$

Negative gradients of the implemented loss functions

Recall that the loss function corresponds to the negative log-likelihood, i.e. $\omega_i = -\ell_i$. The negative gradients of the loss with respect to the additive predictors are then the first partial derivatives of the log-likelihood with respect to the additive predictors, i.e. $-\partial\omega/\partial\eta_{ik}^{(\bullet)} = -\partial(-\ell_i)/\partial\eta_{ik}^{(\bullet)} = \partial\ell_i/\partial\eta_{ik}^{(\bullet)}$, with $\bullet \in \{1, 2, c\}$.

Bivariate binary responses The log-likelihood function for bivariate binary responses is given by:

$$\begin{aligned}\ell_i = {} & y_{i1}y_{i2}\log(p_i^{11}) + y_{i1}(1-y_{i2})\log(p_i^{1(1)} - p_i^{11}) + \\ & (1-y_{i1})y_{i2}\log(p_i^{1(2)} - p_i^{11}) + (1-y_{i1})(1-y_{i2})\log(1 - p_i^{1(1)} - p_i^{1(2)} + p_i^{11}).\end{aligned}$$

where $p_i^{11} = C[\, p_i^{1(1)},\ p_i^{1(2)}; \vartheta_i^{(c)}\,]$, $p_i^{1(1)} = P(Y_{i1} = 1; \vartheta_i^{(1)})$, and $p_i^{1(2)} = P(Y_{i2} = 1; \vartheta_i^{(2)})$. The negative gradients are then:

$$\begin{aligned}\frac{\partial\ell_i}{\partial\eta_i^{(1)}} = \Bigg\{ & \frac{y_{i1}y_{i2}}{p_i^{11}}\frac{\partial p_i^{11}}{\partial p_i^{1(1)}} + \frac{y_{i1}(1-y_{i2})}{p_i^{1(1)} - p_i^{11}}\left[1 - \frac{\partial p_i^{11}}{\partial p_i^{1(1)}}\right] - \\ & \frac{(1-y_{i1})y_{i2}}{p_i^{1(2)} - p_i^{11}}\frac{\partial p_i^{11}}{\partial p_i^{1(1)}} - \frac{(1-y_{i1})(1-y_{i2})}{1 - p_i^{1(1)} - p_i^{1(2)} + p_i^{11}}\left[1 - \frac{\partial p_i^{11}}{\partial p_i^{1(1)}}\right]\Bigg\}\frac{\partial p_i^{1(1)}}{\partial\eta_i^{(1)}},\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell_i}{\partial\eta_i^{(2)}} = \Bigg\{ & \frac{y_{i1}y_{i2}}{p_i^{11}}\frac{\partial p_i^{11}}{\partial p_i^{1(2)}} - \frac{y_{i1}(1-y_{i2})}{p_i^{1(1)} - p_i^{11}}\frac{\partial p_i^{11}}{\partial p_i^{1(2)}} + \\ & \frac{(1-y_{i1})y_{i2}}{p_i^{1(2)} - p_i^{11}}\left[1 - \frac{\partial p_i^{11}}{\partial p_i^{1(2)}}\right] - \frac{(1-y_{i1})(1-y_{i2})}{1 - p_i^{1(1)} - p_i^{1(2)} + p_i^{11}}\left[1 - \frac{\partial p_i^{11}}{\partial p_i^{1(2)}}\right]\Bigg\}\frac{\partial p_i^{1(2)}}{\partial\eta_i^{(2)}},\end{aligned}$$

$$\begin{aligned}\frac{\partial\ell_i}{\partial\eta_i^{(c)}} = \Bigg\{ & \frac{y_{i1}y_{i2}}{p_i^{11}}\frac{\partial p_i^{11}}{\partial\vartheta_i^{(c)}} - \frac{y_{i1}(1-y_{i2})}{p_i^{1(1)} - p_i^{11}}\frac{\partial p_i^{11}}{\partial\vartheta_i^{(c)}} - \\ & \frac{(1-y_{i1})y_{i2}}{p_i^{1(2)} - p_i^{11}}\frac{\partial p_i^{11}}{\partial\vartheta_i^{(c)}} + \frac{(1-y_{i1})(1-y_{i2})}{1 - p_i^{1(1)} - p_i^{1(2)} + p_i^{11}}\frac{\partial p_i^{11}}{\partial\vartheta_i^{(c)}}\Bigg\}\frac{\partial\vartheta_i^{(c)}}{\partial\eta_i^{(c)}}.\end{aligned}$$

Bivariate discrete responses The log-likelihood function for bivariate discrete responses is given by:

$$\begin{aligned}\ell_i &= \log\Big\{C[F_{1i},\ F_{2i};\ \vartheta_i^{(c)}]\ - C[F_{1i}-f_{1i},\ F_{2i};\ \vartheta_i^{(c)})]\ - \\ &\qquad C[F_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{c}]\ + C[F_{1i}-f_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{c}]\Big\} \\ &= \log\Big\{\Delta_i\Big\},\end{aligned}$$

where the marginal CDFs and PDFs have been abbreviated as $F_{ji} = F_j(y_{ij};\boldsymbol{\vartheta}_i^{(j)})$ and $f_{ji} = f_j(y_{ij};\boldsymbol{\vartheta}_i^{(j)})$, $j = 1, 2$, in order to avoid clutter in the notation. The negative gradients are then:

$$\begin{aligned}\frac{\partial \ell_i}{\partial \eta_{ik}^{(1)}} = \frac{1}{\Delta_i}\Bigg\{ &\left[\frac{\partial C[F_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial F_{1i}} - \frac{\partial C[F_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial F_{1i}}\right]\frac{\partial F_{1i}}{\partial \eta_{ik}^{(1)}}\ + \\ &\left[\frac{\partial C[F_{1i}-f_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial F_{1i}-f_{1i}} - \frac{\partial C[F_{1i}-f_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial F_{1i}-f_{1i}}\right]\left[\frac{\partial F_{1i}}{\partial \eta_{ik}^{(1)}} - \frac{\partial f_{1i}}{\partial \eta_{ik}^{(1)}}\right]\Bigg\},\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \eta_{ik}^{(2)}} = \frac{1}{\Delta_i}\Bigg\{ &\left[\frac{\partial C[F_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial F_{2i}} - \frac{\partial C[F_{1i}-f_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial F_{2i}}\right]\frac{\partial F_{2i}}{\partial \eta_{ik}^{(2)}}\ + \\ &\left[\frac{\partial C[F_{1i}-f_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial F_{2i}-f_{2i}} - \frac{\partial C[F_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial F_{2i}-f_{2i}}\right]\left(\frac{\partial F_{2i}}{\partial \eta_{ik}^{(2)}} - \frac{\partial f_{2i}}{\partial \eta_{ik}^{(2)}}\right)\Bigg\},\end{aligned}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \eta_{i}^{(c)}} = \frac{1}{\Delta_i}\Bigg\{ &\frac{\partial C[F_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial \vartheta_i^{(c)}} - \frac{\partial C[F_{1i}-f_{1i},\ F_{2i};\ \vartheta_i^{(c)}]}{\partial \vartheta_i^{(c)}}\ - \\ &\frac{\partial C[F_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial \vartheta_i^{(c)}} + \frac{\partial C[F_{1i}-f_{1i},\ F_{2i}-f_{2i};\ \vartheta_i^{(c)}]}{\partial \vartheta_i^{(c)}}\Bigg\}\frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}},\end{aligned}$$

where $\partial F_{\bullet i}/\partial \eta_{ik}^{(\bullet)} = (\partial F_{\bullet i}/\partial \vartheta_{ik}^{(\bullet)})(\partial \vartheta_{ik}^{(\bullet)}/\partial \eta_{ik}^{(\bullet)})$ and $\partial f_{\bullet i}/\partial \eta_{ik}^{(\bullet)} = (\partial f_{\bullet i}/\partial \vartheta_{ik}^{(\bullet)})(\partial \vartheta_{ik}^{(\bullet)}/\partial \eta_{ik}^{(\bullet)})$, with $\bullet \in \{1, 2\}$.

Bivariate mixed binary-continuous responses The log-likelihood function for mixed binary-continuous responses is the following:

$$\ell_i = (1-y_{i1})\log\left\{\frac{\partial C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{2i}}\right\}\ +\ y_{i1}\log\left\{1 - \frac{\partial C[F_{1i},\ F_{2i};\vartheta_i^{(c)}]}{\partial F_{2i}}\right\} + \log\left[f_{2i}\right],$$

where the marginal CDFs and the marginal PDF of the second margin have been abbreviated as $F_{1i} = F_1(0;\vartheta_i^{(1)})$, $F_{2i} = F_2(y_{i2};\boldsymbol{\vartheta}_i^{(2)})$, and $f_{2i} = f_2(y_{i2};\boldsymbol{\vartheta}_i^{(2)})$, respectively. The negative

gradients are then:

$$\frac{\partial \ell_i}{\partial \eta_i^{(1)}} = \left\{ \frac{(1-y_{i1})}{\frac{\partial C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{2i}}} - \frac{y_{i1}}{1-\frac{\partial C[F_{1i},\ F_{2i};\vartheta_i^{(c)}]}{\partial F_{2i}}} \right\} \frac{\partial^2 C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{1i}\ \partial F_{2i}} \frac{\partial F_{1i}}{\partial \eta_i^{(1)}},$$

$$\frac{\partial \ell_i}{\partial \eta_{ik}^{(2)}} = \left\{ \frac{(1-y_{i1})}{\frac{\partial C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{2i}}} - \frac{y_{i1}}{1-\frac{\partial C[F_{1i},\ F_{2i};\vartheta_i^{(c)}]}{\partial F_{2i}}} \right\} \frac{\partial^2 C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{2i}^2} \frac{\partial F_{2i}}{\partial \eta_{ik}^{(2)}} + \frac{1}{f_{2i}} \frac{\partial f_{2i}}{\partial \eta_{ik}^{(2)}},$$

$$\frac{\partial \ell_i}{\partial \eta_i^{(c)}} = \left\{ \frac{(1-y_{i1})}{\frac{\partial C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial F_{2i}}} - \frac{y_{i1}}{1-\frac{\partial C[F_{1i},\ F_{2i};\vartheta_i^{(c)}]}{\partial F_{2i}}} \right\} \frac{\partial C[F_{1i},\ F_2;\vartheta_i^{(c)}]}{\partial \vartheta_i^{(c)}} \frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}},$$

where $\partial F_{2i}/\partial \eta_{ik}^{(2)} = (\partial F_{2i}/\partial \vartheta_{ik}^{(2)})(\partial \vartheta_{ik}^{(2)}/\partial \eta_{ik}^{(2)})$ and $\partial f_{2i}/\partial \eta_{ik}^{(2)} = (\partial f_{2i}/\partial \vartheta_{ik}^{(2)})(\partial \vartheta_{ik}^{(2)}/\partial \eta_{ik}^{(2)})$.

Part B

In this supplement we provide further details and results of the simulation studies conducted for the bivariate binary (Subsection B1), bivariate discrete (Subsection B2), and bivariate mixed binary-continuous responses (Subsection B3), as well as copula selection (Subsection B4). In the simulation studies we use multivariate proper scoring rules to evaluate the fit of the copula and univariate models. Specifically, we consider the negative log-likelihood (log-score) and the energy score, with the latter being defined as follows. Let $\boldsymbol{y} = (y_1, \ldots, y_d)^\top \in \mathbb{R}^d$ be a new observation with unknown distribution F_Y, and let $\hat{F}_Y$ be a forecast distribution for F_Y. The energy score is then given by

$$ES(F, \boldsymbol{y}) = \frac{1}{n}\sum_{i=1}^{n} ||\boldsymbol{X}_i - \boldsymbol{y}|| - \frac{1}{2n^2}\sum_{i=1}^{n}\sum_{j=1}^{n} ||\boldsymbol{X}_i - \boldsymbol{X}_j||,$$

where $||\cdot||$ denotes the Euclidean norm on $\mathbb{R}^d$ and $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$ are iid copies with distribution $\hat{F}_Y$, for $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{id},) \in \mathbb{R}^d$, $i = 1, \ldots, n$ (Gneiting et al., 2008).

B1 Bivariate binary responses

Data generation We consider three data generating processes (DGPs) with increasing number of noise variables. Specifically, we generate $p_1 = 10$, $p_2 = 100$ and lastly $p_3 = 1000$ covariates, of which only six have are truly informative in one or several of the distribution parameters. These configurations lead to 50% ($p1$), 5% ($p2$) and 0.5% ($p3$) of the covariates being informative, respectively. The bivariate distribution of the binary components is created using a Gaussian copula with varying correlation between the margins. On average, the dependence between the margins of the synthetic data in terms of Kendall's τ lies within $[-0.993;\ 0.993]$, i.e. it ranges between very strong negative to very strong positive dependence. We generate the first margin from a probit model and the second margin from a cloglog model. Thus, the model has $K = 3$ distribution parameters and the DGP with p_3 covariates represents a high-dimensional setting with $p_3 \gg n$, resulting in effectively $p_3 \times K = 3000$ covariates since we fit all regressors to each distribution parameter. For this scenario we only consider DGPs with linear effects of the covariates, which reflect the data analysed in Section 4.1. Following Strömer et al. (2023), we generate the p_q, $q = 1, 2, 3$ covariates by sampling from a multivariate Gaussian distribution with Toeplitz covariance structure of the form $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p_q$, with $\rho = 0.5$ denoting the correlation between consecutive covariates x_j and x_{j+1}. We consider the following linear predictors

$$\begin{aligned}
\Phi^{-1}\left(p_i^{1(1)}\right) = \eta_{i1}^{(1)} &= -1x_{i2} + 0.5x_{i3} + 1x_{i4} - 0.5x_{i6}, \\
\log\left(\left(-\log(1 - p_i^{1(2)}\right)\right) = \eta_{i1}^{(2)} &= 0.5x_{i1} - 1x_{i2} + 0.75x_{i3}, \\
\tanh\left(\vartheta_i^{(c)}\right)^{-1} = \eta_i^{(c)} &= 0.5x_{i2} - 1.5x_{i3} + 1.5x_{i4}.
\end{aligned}$$

Results Table B2, Column (1) summarizes the performance scores. Overall, the copula model exhibits a better performance in terms of the negative log-likelihood (log-score) as well as the energy score, indicating a better fit of the bivariate distribution. In terms of univariate scores (Brier score and AUC), both the univariate and copula models show similar results, with the copula model outperforming the univariate model at predicting the second margin in the high-dimensional setting ($p_3 = 1000$). The selection rates of informative and non-informative covariates in each of the distribution parameters are given in Table B3, Column (1). These are the percentage of instances in which the informative and non-informative variables entered the model's additive predictors, respectively. Overall, the selection rates rates of non-informative covariates in the distribution parameters belonging to the marginal distributions are slightly higher as compared to corresponding rates obtained from separate univariate models. This is somewhat expected given the increased model complexity. On the other hand, the differences decrease considerably as the number of non-informative covariates increases. The selection rates of non-informative covariates in the dependence parameter $\vartheta^{(c)}$ are considerably lower than that of effects in the parameters of the margins, whereas the selection rates of informative variables are slightly lower than 100%. This suggests that the shrinkage is strongest in the dependence parameter and the chosen criterion to determine $\texttt{m}_{\texttt{stop}}$ slightly underfits the effects in the dependence. However, both the copula and univariate models correctly select the informative covariates in all margins. Figure B1(a) depicts the estimated coefficients with each row corresponding to $p_1 = 10$, $p_2 = 100$, $p_3 = 1000$, respectively. In low-dimensional settings (p_1), the copula model matches the univariate models in producing accurate estimates of all linear coefficients in the marginal distributions. In high-dimensional settings (p_3), the estimated coefficients corresponding to the margins exhibit similar performance as in settings with p_1 and p_2.

B2 Bivariate discrete responses

Data generation We consider two DGPs for bivariate count responses each including $p = 10$ covariates. In the first DGP, the covariates have a strictly linear effect on the distribution parameters, whereas in the second DGP we consider non-linear effects. These configurations lead to 60% and 50% of the covariates being informative in the linear and non-linear DGPs, respectively. The bivariate discrete distribution is constructed using a combination of a Zero-Altered Logarithmic distribution (ZALG, margin 1) with two parameters, and a Zero-Inflated Negative Binomial Type I distribution (ZINBI, margin 2), which has three parameters. The marginal distributions and number of covariates are chosen to resemble the data studied in Section 4.2. The components are linked through a Joe copula, which allows to model positive dependence as well as upper tail dependence between the margins. The additive predictor of the dependence parameter $\vartheta_i^{(c)}$ covers Kendall's τ values within [0.275; 0.899], ranging from moderate to very strong positive dependence between Y_1 and Y_2. The covariates are sampled from independent univariate Uniform distributions with support between 0 and 1,

i.e. $X_r \sim U[0,1],\ \forall r = 1, \ldots, 10$. Overall, the bivariate distribution consists of six parameters with the following additive predictors

Linear DGP:	Non-linear DGP:
$\log\left(\frac{\vartheta_{i1}^{(1)}}{1-\vartheta_{i1}^{(1)}}\right) = -1x_{i1} + 1x_{i3},$	$\log\left(\frac{\vartheta_{i1}^{(1)}}{1-\vartheta_{i1}^{(1)}}\right) = \frac{1}{2}\left(x_{i1}^{3/2} - 2\cos(3x_{i1})\right),$
$\log\left(\frac{\vartheta_{i2}^{(1)}}{1-\vartheta_{i2}^{(1)}}\right) = +1x_{i4} + 1x_{i5} - 2x_{i8},$	$\log\left(\frac{\vartheta_{i2}^{(1)}}{1-\vartheta_{i2}^{(1)}}\right) = -80\left(x_{i3}^{3/2} - x_{i3}^{4/3}\right),$
$\log\left(\vartheta_{i1}^{(2)}\right) = +1.5x_{i1} - 1.5x_{i2},$	$\log\left(\vartheta_{i1}^{(2)}\right) = -0.7\exp\left(x_{i2}^2\right) + \exp\left(x_{i2}^{0.4}\right),$
$\log\left(\vartheta_{i2}^{(2)}\right) = -0.75x_{i2} + 1x_{i4},$	$\log\left(\vartheta_{i2}^{(2)}\right) = 3 - 1.5\left(1.5\cos(2x_{i5}) + 3\tanh(x_{i5})\right),$
$\log\left(\frac{\vartheta_{i3}^{(2)}}{1-\vartheta_{i3}^{(2)}}\right) = -0.75x_{i2} + 1x_{i3},$	$\log\left(\frac{\vartheta_{i3}^{(2)}}{1-\vartheta_{i3}^{(2)}}\right) = -3 - 0.7\left(\sin(x_{i1}) - \exp(x_{i1})^2\right),$
$\log\left(\vartheta_i^{(c)} - 1\right) = -0.5x_{i2} + 1.5x_{i3} + 1.5x_{i5},$	$\log\left(\vartheta_i^{(c)} - 1\right) = 2\sin(4x_{i4}).$

Note that in case of the linear DGP, seven out of the ten covariates have a non-zero effect on the distribution parameters with five of those overlapping and one having an effect uniquely on one parameter. In the non-linear DGP, six covariates are informative and once again there is some overlap in the informative covariates across parameters.

Results The performance metrics for the bivariate count response scenario are summarized in Table B2, Column (2). In terms of log- and energy scores, our proposed copula approach outperforms the univariate models considerably. The copula also leads to a smaller MSEP for predicting Y_2, whereas the univariate model for Y_1 outperforms the copula in terms of MSEP. The selection rates in Table B3, Column (2) demonstrate that the copula model selects more non-informative variables in the first margin, as well as in the first parameter of margin 2 compared to the univariate models in the linear DGP. However, the selection rates of informative covariates from the univariate models are considerably lower in the other two parameters of margin 2 compared to those of our copula model. These once again point out that the dependence parameter experiences the strongest shrinkage of the covariate effects. In the non-linear DGP, the selection rates of non-informative covariates are similar for the copula and univariate models in the first margin. In contrast, in the second margin the copula model tends to select too many non-informative covariates in the first parameter of margin 2. Concerning the linear effects, our approach performs well given the overlap between covariate effects and distribution parameters. The shrinkage on the estimated coefficients corresponding to the dependence parameter exhibits a very similar behaviour to that observed in the bivariate binary scenario with $p_1 = 10$. The univariate models tend to underestimate the covariate effects in the second parameter of margin 1 and the third parameter of margin 2. Furthermore, the univariate models display a slightly higher variance in the estimated coefficients compared to those derived from the copula models. This observation is supported by Figure B2(b), which

indicates selection rates of informative covariates in the non-linear DGP are consistently at 100% across all models.

B3 Bivariate mixed responses

Data generation For the mixed binary-continuous response scenario we generate the binary margin using a probit model, whereas the continuous margin follows a heteroskedastic Gaussian distribution. The components are linked through a Clayton copula rotated by 270°, which supports dependence between very high values of Y_1 and low values of Y_2. The choice of margins and copula is based on the data on children undernutrition analysed in Subsection 4.3. A total of $p = 10$ covariates are obtained from independent univariate Uniform distributions between 0 and 1, i.e. $X_r \sim U[0,1],\ \forall r = 1,\ldots,10$. Once again we study both a DGP with only linear effects and another with non-linear effects of the covariates. In these configurations 50% (linear DGP) and 30% (non-linear DGP) of the covariates are informative. The bivariate distribution features four parameters with following additive predictors

Linear DGP:

$$\begin{aligned}
\Phi^{-1}\left(\vartheta_{i1}^{(1)}\right) &= 1.5x_{i2} - 1x_{i3} + 1.5x_{i4},\\
\vartheta_{i1}^{(2)} &= 0.5x_{i2} + 1.5x_{i3},\\
\log\left(\vartheta_{i2}^{(2)}\right) &= 1x_{i5},\\
\log\left(-\vartheta_{i}^{(c)}\right) &= 1.5x_{i5} - 1.5x_{i6},
\end{aligned}$$

Non-linear DGP:

$$\begin{aligned}
\Phi^{-1}\left(\vartheta_{i1}^{(1)}\right) &= \frac{1}{2}\left(x_{i1}^{3/2} - 2\cos(3x_{i1})\right),\\
\vartheta_{i1}^{(2)} &= -0.7\exp\left(x_{i1}^{2}\right) + \exp\left(x_{i1}^{0.4}\right),\\
\log\left(\vartheta_{i2}^{(2)}\right) &= -0.5 + \cos(2x_{i2})\\
\log\left(-\vartheta_{i}^{(c)}\right) &= -1 + 3\sin(4x_{i3}).
\end{aligned}$$

In the linear DGP, only five covariates have a non-zero effect on the distribution parameters and once again there is some overlap between informative covariates and distribution parameters. In the non-linear DGP, there are four informative covariates with some overlap between parameters and informative covariates.

Results From Table B2, Column (3) it can be observed that the copula model outperforms the univariate models in terms of the log and energy scores, with the difference in these two models becoming more apparent in the non-linear DGP. In the linear DGP, the copula model performs better than the univariate models in terms of the log-score but slightly worse in terms of the energy score, albeit the difference in energy scores is 0.001 and the univariate models exhibit a much larger standard deviation in the aforementioned score. Regarding the univariate scores, once again both copula and univariate models exhibit similar performance. In the non-linear DGP, both models are able to recover the true non-linear effects of the informative covariates (see Figure B2(c)). In the linear DGP a slight improvement in efficiency can be observed from the copula model. We remark that the boosted copula model once again exhibits a higher degree of shrinkage of the effects present in the copula parameter, as indicated by the selection rates in Table B3, Column (3). Similar to the other two response

scenarios, both copula and univariate models effectively identify the informative covariates across all distribution parameters of the margins. In the non-linear DGP, the copula model demonstrates a tendency to exhibit higher selection rates of non-informative variables within the continuous margin, while remaining competitive in the binary margin.

B4 Copula selection via the out-of-sample negative log-likelihood

We investigate the performance of the the out-of-sample negative log-likelihood to select the correct copula function under different a growing number of candidate covariates that enter the model. We generate bivariate binary data from a Gaussian copula with varying dependence using the configuration from Subsection B1. Hence the predictors that generate the dependent bivariate binary responses are given by:

$$\begin{aligned}
\Phi^{-1}\left(p_i^{1(1)}\right) = \eta_{i1}^{(1)} &= -1x_{i2} + 0.5x_{i3} + 1x_{i4} - 0.5x_{i6}, \\
\log\left(\left(-\log(1-p_i^{1(2)}\right)\right) = \eta_{i1}^{(2)} &= 0.5x_{i1} - 1x_{i2} + 0.75x_{i3}, \\
\tanh\left(\vartheta_i^{(c)}\right)^{-1} = \eta_i^{(c)} &= 0.5x_{i2} - 1.5x_{i3} + 1.5x_{i4}.
\end{aligned}$$

We then fit the following candidate distributions to the data:

1. Independent Bernoulli margins.
2. Gaussian copula with Bernoulli margins (correct specification).
3. Clayton copula with Bernoulli margins.
4. Clayton copula rotated by 90° with Bernoulli margins.
5. Clayton copula rotated by 180° with Bernoulli margins.
6. Gumbel copula with Bernoulli margins.

The link functions of the margins are correctly specified, and only the copula function is subject to misspecification. We consider $p_1 = 10$, $p_2 = 100$, and $p_3 = 1000$ covariates in the data. Similar to Subsection B1, 50% ($p1$), 5% ($p2$) and 0.5% ($p3$) of the covariates are informative in the respective cases. After finding the optimal number of fitting iteration of each boosting model as described in Algorithm 1, the out-of-sample negative log-likelihood is computed on a test set of size $n_{test} = 1000$ observations.

Results Table B1 shows the values of the out-of-sample negative log-likelihood for all candidate models and levels of sparsity. Overall, the correct copula (column 2, Gaussian) performs best.

Table B1: Simulation study. Out-of-sample negative log-likelihood (standard errors shown in parentheses) of fits from six candidate distributions (columns) to the true generating process that employs a Gaussian copula. The rows correspond to the different levels of sparsity. Lower values indicate better perfomance.

	True copula: Gaussian											
	Independence		Gaussian		Clayton 0°		Clayton 90°		Clayton 180°		Gumbel 0°	
$p_1 = 10$	974.33	(23.03)	**881.83**	(22.00)	916.67	(23.64)	1221.81	(24.70)	1324.32	(13.77)	921.96	(23.65)
$p_2 = 100$	985.59	(21.28)	**897.97**	(21.15)	953.82	(24.72)	1239.07	(20.92)	1327.56	(13.04)	961.56	(24.26)
$p_3 = 1000$	998.88	(19.70)	**915.30**	(20.17)	987.77	(22.62)	1256.22	(22.84)	1329.93	(13.93)	1004.60	(22.61)

Table B2: Simulation study. Performance metrics for the simulation studies for the copula (C) and univariate models (U), $\star$ identifies the non-linear DGP. Values are mean scores from 200 independent replicates (each evaluated on the test dataset), whereas parentheses show the respective standard deviations.

			(1)		(2)	(3)
Score	Model		Bivariate binary		Bivariate count	Mixed
		$p_1 = 10$	$p_2 = 100$	$p_3 = 1000$	$p = 10$	$p = 10$
Log	C	887.186 (24.103)	914.295 (22.364)	954.294 (21.120)	1579.591 (48.609)	1990.636 (28.791)
	U	956.639 (24.892)	971.851 (22.868)	998.185 (20.904)	1961.514 (64.262)	2011.168 (27.048)
	$C\star$	-	-	-	2442.732 (57.129)	1755.638 (31.690)
	$U\star$	-	-	-	2822.878 (64.924)	1911.316 (31.211)
Energy	C	0.279 (0.008)	0.284 (0.007)	0.293 (0.007)	0.725 (0.041)	0.687 (0.013)
	U	0.285 (0.008)	0.290 (0.007)	0.298 (0.007)	0.741 (0.041)	0.710 (0.466)
	$C\star$	-	-	-	1.581 (0.071)	0.673 (0.015)
	$U\star$	-	-	.	1.601 (0.069)	0.679 (0.015)
Brier (Y_1)	C	0.142 (0.006)	0.144 (0.007)	0.148 (0.006)	-	0.200 (0.006)
	U	0.142 (0.006)	0.144 (0.007)	0.148 (0.006)	-	0.199 (0.006)
	$C\star$	-	-	-	-	0.174 (0.006)
	$U\star$	-	-	-	-	0.174 (0.006)
Brier (Y_2)	C	0.177 (0.006)	0.180 (0.006)	0.185 (0.006)	-	-
	U	0.177 (0.006)	0.180 (0.006)	0.185 (0.005)	-	-
	$C\star$	-	-	-	-	-
	$U\star$	-	-	-	-	-
AUC (Y_1)	C	0.879 (0.010)	0.876 (0.012)	0.870 (0.012)	-	0.760 (0.015)
	U	0.879 (0.010)	0.876 (0.012)	0.871 (0.012)	-	0.760 (0.015)
	$C\star$	-	-	-	-	0.816 (0.012)
	$U\star$	-	-	-	-	0.816 (0.012)
AUC (Y_2)	C	0.796 (0.015)	0.791 (0.014)	0.776 (0.015)	-	-
	U	0.796 (0.015)	0.791 (0.014)	0.780 (0.015)	-	-
	$C\star$	-	-	-	-	-
	$U\star$	-	-	-	-	-
MSEP (Y_1)	C	-	-	-	1.084 (0.133)	-
	U	-	-	-	1.073 (0.132)	-
	$C\star$	-	-	-	1.563 (0.302)	-
	$U\star$	-	-	-	1.556 (0.302)	-
MSEP (Y_2)	C	-	-	-	2.504 (0.557)	1.190 (0.058)
	U	-	-	-	2.412 (0.581)	1.188 (0.058)
	$C\star$	-	-	-	10.799 (1.150)	1.293 (0.072)
	$U\star$	-	-	-	11.058 (1.221)	1.290 (0.072)
Copula			Gaussian		Joe	Rotated Clayton 270°
Kendall's τ range			$[-0.993;\ 0.993]$		$[0.275;\ 0.899]$	$[-0.787;\ -0.019]$

Gradients stabilised using L_2 norm, step-length $\mathtt{s_{step}} = 0.1$. $n_{\text{train}} = 1000$, $n_{\text{test}} = 1000$, $n_{\mathtt{mstop}} = 1500$.

Table B3: Simulation study. Selection rates (in %) of informative (x_{inf}) and non-informative ($x_{\text{n-inf}}$) covariates for the copula (C) and univariate models (U) for each distribution parameter, $\star$ denotes non-linear DGP. Values are averages over 200 independent datasets.

	(1) Bivariate binary						(2) Bivariate count		(3) Binary & continuous	
	$p_1 = 10$		$p_2 = 100$		$p_3 = 1000$		$p = 10$		$p = 10$	
	x_{inf}	$x_{\text{n-inf}}$	x_{inf}	$x_{\text{n-inf}}$	x_{inf}	$x_{\text{n-inf}}$	x_{inf}	$x_{\text{n-inf}}$	x_{inf}	$x_{\text{n-inf}}$
Linear DGP										
Copula model (C)										
$\vartheta_1^{(1)}$	100	78.833	100	26.297	100	6.189	99	59.062	100	64.000
$\vartheta_2^{(1)}$	-	-	-	-	-	-	100	64.714	-	-
$\vartheta_1^{(2)}$	100	77.929	100	28.387	100	7.142	100	72.812	100	74.500
$\vartheta_2^{(2)}$	-	-	-	-	-	-	97	50.250	100	81.611
$\vartheta_3^{(2)}$	-	-	-	-	-	-	75.750	48.812	-	-
$\vartheta^{(c)}$	95.125	54.167	70.000	6.609	55.750	0.209	89.500	40.643	87.500	27.687
Univariate models (U)										
$\vartheta_1^{(1)}$	100	69.583	100	22.745	100	4.785	95.250	38.250	100	56.071
$\vartheta_2^{(1)}$	-	-	-	-	-	-	100	46.571	-	-
$\vartheta_1^{(2)}$	100	70.643	100	20.722	100	4.643	100	60	100	32.187
$\vartheta_2^{(2)}$	-	-	-	-	-	-	90.750	56.750	100	38.778
$\vartheta_3^{(2)}$	-	-	-	-	-	-	36	18.375	-	-
Non-linear DGP										
Copula model ($C\star$)										
$\vartheta_1^{(1)}$	-	-	-	-	-	-	100	29.944	100	33.111
$\vartheta_2^{(1)}$	-	-	-	-	-	-	100	44.444	-	-
$\vartheta_1^{(2)}$	-	-	-	-	-	-	100	99.500	100	83.000
$\vartheta_2^{(2)}$	-	-	-	-	-	-	100	34.278	100	83.611
$\vartheta_3^{(2)}$	-	-	-	-	-	-	100	4.889	-	-
$\vartheta^{(c)}$	-	-	-	-	-	-	100	18.944	100	0.278
Univariate models ($U\star$)										
$\vartheta_1^{(1)}$	-	-	-	-	-	-	100	27.556	100	38.278
$\vartheta_2^{(1)}$	-	-	-	-	-	-	100	22.222	-	-
$\vartheta_1^{(2)}$	-	-	-	-	-	-	99.500	59.778	100	25.722
$\vartheta_2^{(2)}$	-	-	-	-	-	-	98.500	53.278	100	36.611
$\vartheta_3^{(2)}$	-	-	-	-	-	-	100	8.889	-	-

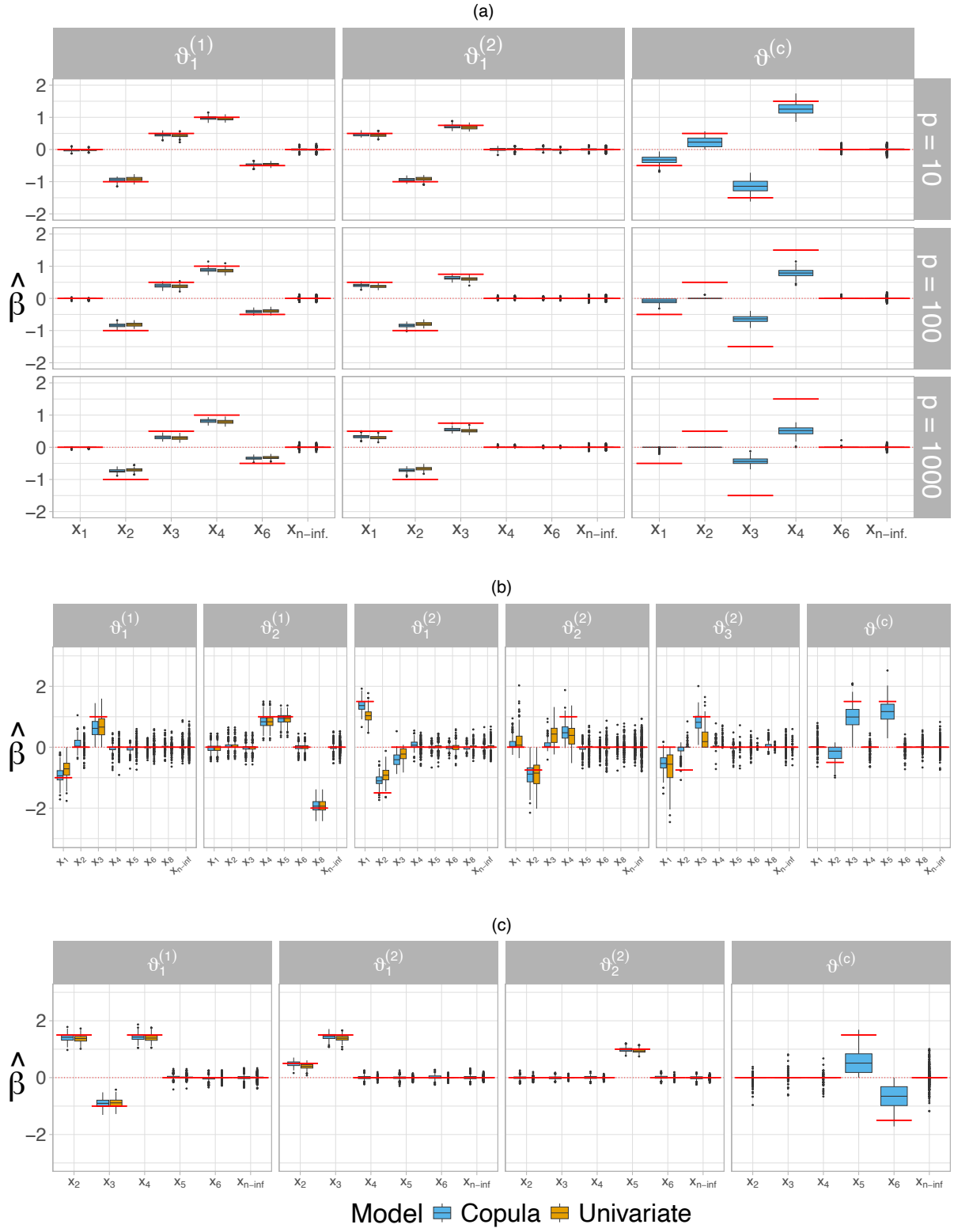

Figure B1: Simulation study. Estimated coefficients of informative covariates $x_1, \ldots, x_6$ and non-informative ($x_{\text{n-inf.}}$) covariates from copula and univariate models across distribution parameters for linear DGPs of Simulation B1 (a, Gaussian copula), B2 (b, Joe copula), B3 (c, rotated Clayton copula by 270°). Results obtained using 200 independent datasets.

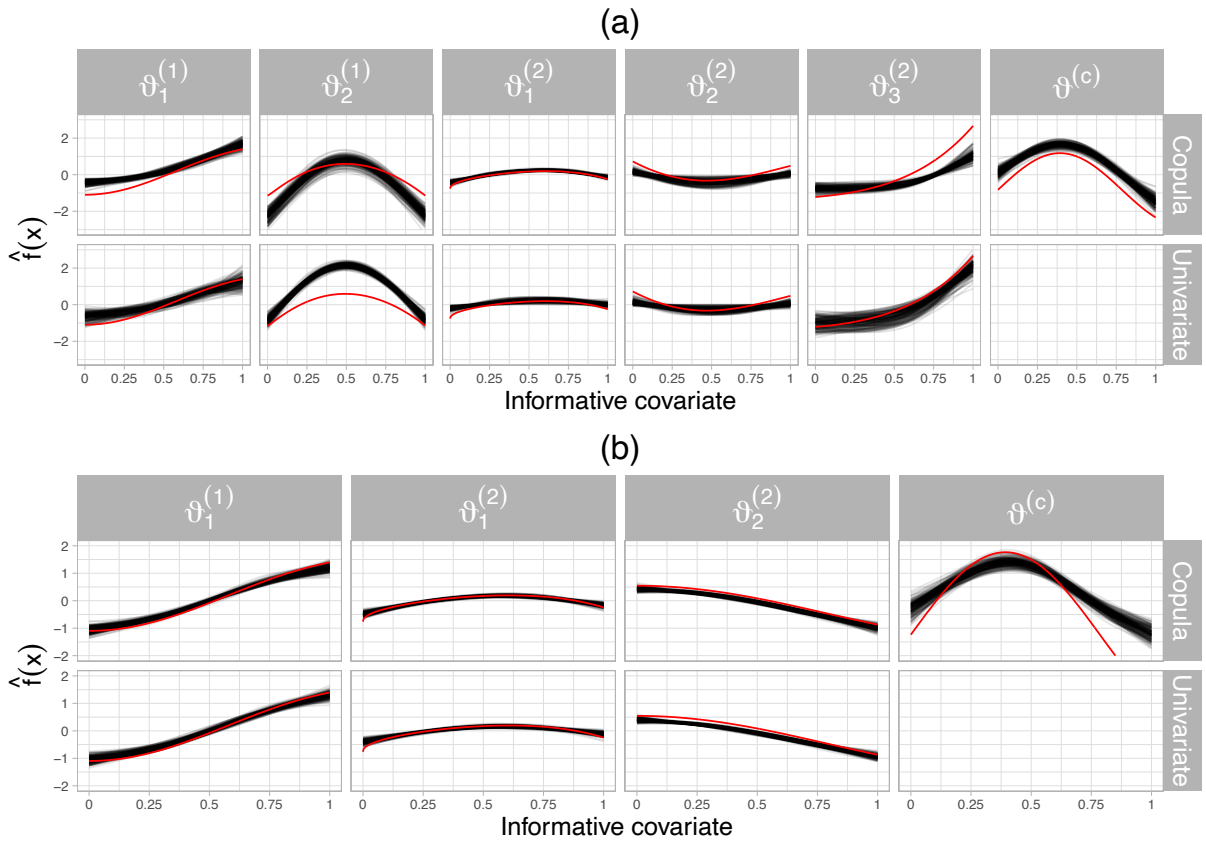

Figure B2: Simulation study. Estimated effects of the informative covariates from copula and univariate models across distribution parameters for non-linear DGPs of Simulation B2 (a, Joe copula) and B3 (b, rotated Clayton copula by 270°). Red lines indicate true effects. Results obtained using 200 independent datasets. The x-axis shows the corresponding covariate x_j and the y-axis the functional estimates $\hat{f}_j(x_j)$.

Part C

C1 Supporting results of Application 4.2

Table C1: Application 4.2. Out-of-sample negative log-likelihoods of the candidate discrete marginal distributions evaluated on $n_{test} = 778$ observations.

Marginal distribution	Negative log-likelihood	
	Margin 1 (`doctorco`)	Margin 2 (`prescrib`)
Poisson	537.250	960.914
Geometric	523.207	906.440
Negative Binomial (I)	527.338	900.263
Zero-Altered Logarithmic	**522.664**	912.034
Zero-Altered Negative Binomial	524.863	898.757
Zero-Inflated Poisson	531.583	911.166
Zero-Inflated Negative Binomial	528.512	**895.975**

Table C2: Application 4.2. Out-of-sample negative log-likelihoods of the candidate copula functions and the bivariate Poisson distribution evaluated on $n_{test} = 778$ observations.

Copula	Negative log-likelihood
Independence	1418.639
Gaussian	1394.516
Clayton	**1392.381**
Gumbel	1406.620
Joe	1412.192
Farlie-Gumbel-Morgenstern	1401.902
Ali-Mikhail-Haq	1401.150
Bivariate Poisson distribution	1468.074

Table C3: Application 4.2. Selected base-learners across distribution parameters of the joint bivariate distribution of `doctorco` (ZALG) and `prescrib` (ZINBI).

Base-learner	Type	ZALG		ZINBI			Clayton copula
		$\vartheta_1^{(1)}$	$\vartheta_2^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta_3^{(2)}$	$\vartheta^{(c)}$
`gender`	Linear	✓	✓	✓	✓	✓	✓
`age`	Non-linear	✓	✓	✓	✓	✓	✓
`income`	Non-linear	✓	✓	✓	✓		
`age:income`	Interaction			✓			
`age:gender`	Interaction	✓	✓	✓	✓	✓	
`income:gender`	Interaction			✓			

C2 Description and summary statistics of variables featured in Application 4.3

Table C4: Variables of Section 4.3. Responses are fever (binary, row 1) and wasting (continuous, row 2); covariates entering the model non-linearly are child's age, breastfeeding and the mother's body-mass-index (rows 3–5); the binary covariate child's gender enters the model linearly (row 5) and the district in India (row 6) enters the model as a discrete spatial effect.

Variable	Description	Type	Mean (s.d.)
`fever`	Fever experienced within two weeks preceding survey interview	Binary	0.307 (0.461)
`wasting`	Low weight-for-height	Continuous	−79.144 (123.367)
`cage`	Age of the child in months	Continuous	17.255 (10.148)
`breastfeeding`	Months of breastfeeding	Continuous	14.076 (8.751)
`mbmi`	Mother's Body-Mass-Index	Continuous	19.783 (2.937)
`cgender`	Gender of the child (1 female, 0 male)	Binary	0.476
`distH`	District of residence	Factor	-
Number of districts: 438, $n = 24{,}286$.			

References

Gneiting, T., Stanberry, L. I., Grimit, E. P., Held, L., and Johnson, N. A. (2008). Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*, 17(2):211–235.
URL: `https://doi.org/10.1007/s11749-008-0114-x`.

Rigby, R. A., Stasinopoulos, M. D., Heller, G. Z., and De Bastiani, F. (2019). *Distributions for modeling location, scale, and shape: Using GAMLSS in R.* Chapman and Hall/CRC, New York.
URL: `https://doi.org/10.1201/9780429298547`.

Strömer, A., Klein, N., Staerk, C., Klinkhammer, H., and Mayr, A. (2023). Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*, 42(11):1779–1801.
URL: `https://doi.org/10.1002/sim.9699`.

Appendix D: Boosting distributional copula regression for bivariate right-censored time-to-event data (with Supplement)

Joint work with Nadja Klein, Andreas Groll, and Andreas Mayr.

Briseño Sanchez, G., Klein, N., Groll, A., & Mayr, A. (2024). Boosting distributional copula regression for bivariate right-censored time-to-event data. [Manuscript submitted for publication.]

Currently published on arXiv: https://doi.org/10.48550/arXiv.2412.15041.

Boosting distributional copula regression for bivariate right-censored time-to-event data

Guillermo Briseño Sanchez[1], Nadja Klein[1],
Andreas Groll[2] and Andreas Mayr[3]

[1]Methods for Big Data, Scientific Computing Center, Karlsruhe Institute of Technology, Karlsruhe, Germany,

[2]Statistical Methods for Big Data, Department of Statistics, TU Dortmund University, Dortmund, Germany,

[3]Department of Medical Biometry and Statistics, Philipps University of Marburg, Marburg, Germany.

April 28, 2025

Abstract

We propose a highly flexible distributional copula regression model for bivariate time-to-event data in the presence of right-censoring. The joint survival function of the response is constructed using parametric copulas, allowing for a separate specification of the dependence structure between the time-to-event outcome variables and their respective marginal survival distributions. The latter are specified using well-known parametric distributions such as the log-Normal, log-Logistic (proportional odds model), or Weibull (proportional hazards model) distributions. Hence, the marginal univariate event times can be specified as parametric (also known as *Accelerated Failure Time*, AFT) models. Embedding our model into the class of generalized additive models for location, scale and shape, possibly all distribution parameters of the joint survival function can depend on covariates. We develop a component-wise gradient-based boosting algorithm for estimation. This way, our approach is able to conduct data-driven variable selection. To the best of our knowledge, this is the first implementation of multivariate AFT models via distributional copula regression with automatic variable selection via statistical boosting. A special merit of our approach is that it works for high-dimensional ($p \gg n$) settings. We illustrate the practical potential of our method on a high-dimensional application related to semi-competing risks responses in ovarian cancer. All of our methods are implemented in the open source statistical software R as add-on functions of the package gamboostLSS.

Keywords: Accelerated failure time model; Variable selection; Dependence modelling; Semi-competing risks; Survival analysis.

1 Introduction

Advancements in molecular medicine, genetics and digital transformation of healthcare have facilitated the collection of large-scale data structures related to individual patients. Some prominent examples are Genome-Wide Association Studies (GWAS; Uffelmann et al., 2021) and The Cancer Genome Atlas Program (TCGA; TCGA Research Network, 2024). Various techniques have been developed to analyse such "omics" data in a concise, scalable manner, while at the same time preserving the interpretability of the results. An important challenge when facing a vast amount of potentially influencing factors is to find a subset of such factors that has the most impact on the outcome of interest. For exploratory analyses, taking into account the entire information simultaneously instead of performing multiple univariate analyses that ignore the remaining variables in the data is of great importance. Individual analysis of the potential influencing factors without consideration of the remainder could lead to estimation bias or falsely informative selected variables. Therefore, the aforementioned *variable selection* procedure should have as least input from an analyst as possible and instead rely on data-driven techniques.

Compared to classical continuous or binary endpoints, time-to-event data are typically incomplete or *censored* for individual where the event of interest was not observed. Conducting statistical analysis without taking censoring into account leads to bias in the estimation, which could result in incorrect treatment, diagnosis and prognosis. Time-to-event analyses or "survival analyses" (Klein and Moeschberger, 2003) explicitly account for censored observations, see Beis et al. (2024) for a review focused on clinical applications. When analysing univariate censored event-time responses in a regression context, the Cox proportional hazards model (Cox, 1972) is one of the most popular methods, although the interpretation of hazards remains challenging (Heller, 2024; Beyersmann et al., 2024).

A wide range of tools for analysing univariate time-to-event responses accompanied by a large amount of covariate information are available. One commonly used technique to navigate large data structures with high-dimensional covariate information is based on univariate modelling paired with hypothesis testing (Chowdhury and Turin, 2020; Jenssen et al., 2002). That is, the response is modelled as function of one covariate, and after carrying out all of the univariate combinations the p-values obtained from the statistical tests are sorted in ascending order. Afterwards, a subset that includes the "most significant" variables is chosen. In the context of genomics, where gene expression data is overwhelmingly large relative to the number of observations, following the aforementioned approach may lead to poor results (Lo et al., 2015). More sophisticated variable selection approaches such as the LASSO have been adapted to the Cox model (Tibshirani, 1997) as well as Accelerated Failure Time (AFT) or parametric survival models, see e.g. Parsa et al. (2024). More recently, "black-box" or less interpretable methods have also been proposed by Ishwaran et al. (2011), Norman et al. (2024), and Wang

and Li (2017), to name a few, and Salerno and Li (2023) for a review. The main limitation of the aforementioned contributions is their restriction to univariate time-to-event responses.

While a broad literature on multivariate time-to-event analysis exists, variable selection in these models remains somewhat unaddressed. Current proposed approaches do not scale to higher dimensions of covariate information or have not adopted a data-driven approach to variable selection. Marra and Radice (2020) introduced a flexible class of bivariate time-to-event models using parametric copulas. In their approach, the marginal survival functions are modelled semi-parametrically using additive regression techniques and smooth functions of time. Sun and Ding (2019) proposed a copula-based model for time-to-event analysis as well, albeit their implementation is tailored towards interval-censored responses, marginal distributions being of the same family, and the dependence between the event times cannot depend on covariates. A copula-based model for correlated event times was proposed by Emura et al. (2017). However, their approach resorts to "Cox-type" specifications of the marginal survival functions and is also restricted to a constant dependence parameter. Moreover, Emura et al. (2018) extended their proposed model to (indirectly) account for high-dimensional covariates using a "composite covariate" (Tukey, 1993), where a linear combination of coefficients and covariates summarises the high dimensional covariate vector to a scalar variable or index. This new scalar variable is used as proxy for the original high-dimensional covariate information.

In summary, limitations of the currently available methods for time-to-event analysis may be assigned to three categories: (1) The approaches offer solutions for high-dimensional covariates, but are restricted to univariate time-to-event responses. (2) The approaches are able to model multivariate event times, but restrictions exist regarding the flexibility of the marginal survival functions, dependence structure, or covariate effects. (3) The methods are able to handle multivariate responses, but do not scale to high-dimensional covariates or rely on heuristics or non-interpretable techniques to tackle this issue.

We aim to address these gaps by proposing a flexible approach that allows to account for different types of covariate effects in a copula-based multivariate time-to-event model. Furthermore, our proposal allows for scalable, data-driven variable selection via estimation through statistical boosting (Bühlmann and Hothorn, 2007). Boosting has been explored previously in a univariate time-to-event context using different modelling approaches. For example, Binder et al. (2009) applied boosting to high-dimensional competing risks data. He et al. (2016) applied it for false discovery control, whereas Mayr et al. (2016) focused on optimising the concordance index. More recently, Morris et al. (2020) released a package for boosting stratified Cox proportional hazards models. In terms of multivariate responses, Griesbach et al. (2021) proposed a boosting algorithm for variable allocation and selection in the context of joint models for longitudinal and survival data, see Rizopoulos (2012) for more on this model class. Lastly, the alternative modelling paradigm of "first-hitting-time" was combined with

boosting by De Bin and Stikbakke (2023). Our proposed statistical modelling framework allows to construct flexible parametric joint survival functions based on the copula approach. A main advantage is to potentially model all parameters of the joint survival function as functions of covariates using structured additive predictors (Wood, 2017). This in principle gives directly interpretable models. However, because we allow all distribution parameters to depend on covariates, scalable and data-driven variable selection without any input from the analyst is highly desirable. To achieve this goal, we suggest estimation via statistical boosting building on the work of Hans et al. (2023) and Briseño Sanchez et al. (2024). Compared to these authors, we thereby provide boosting methodology and software implementation for distributional copula regression by allowing the responses to be subject to independent right-censoring. To the best of our knowledge, this is the only publicly available software implementation that allows to fit bivariate time-to-event models which combines a wide range of copula functions, marginal distributions, covariate effects and data-driven variable selection.

The remainder of this manuscript is structured as follows: Section 2 presents distributional copula regression for bivariate right-censored time-to-event as well as semi-competing risks responses and outlines our boosting algorithm. Section 3 documents our simulation studies and respective results. In Section 4 we analyse a high-dimensional ($p \gg n$) micro-array dataset related to patients suffering from ovarian cancer in which the time-to-event responses, time of tumour progression and time of death, follow a semi-competing risks data generating process. We model the joint survival function of the time of tumour progression and time of death as a function of genomic as well as clinical information. Additionally, we illustrate the model-building process that involves selecting marginal distributions and the copula function. Lastly, a discussion is given in Section 5.

2 Methods

In this section, we briefly introduce right-censored and semi-competing risks time-to-event responses. Afterwards we outline our distributional copula regression framework for bivariate right-censored time-to-event responses and describe how to perform estimation by means of component-wise gradient boosting.

2.1 Right-censored time-to-event responses

A univariate right-censored time-to-event response is comprised of $Y = \min\{T, T^{cens}\}$ and its censoring indicator $\delta = \mathbb{1}\{T \leq T^{cens}\}$, where T is the true event time and T^{cens} is an independent, random, uninformative censoring time. In addition, we assume that we have some covariate information $\boldsymbol{x}$ available. In what follows, we are concerned with bivariate right-censored time-to-event responses which consist of two univariate right-censored event times $\boldsymbol{Y} = (Y_1, Y_2)^\top$ and their corresponding indicators $\boldsymbol{\delta} = (\delta_1, \delta_2)^\top$, and we write $(\boldsymbol{Y}, \boldsymbol{\delta})$ for

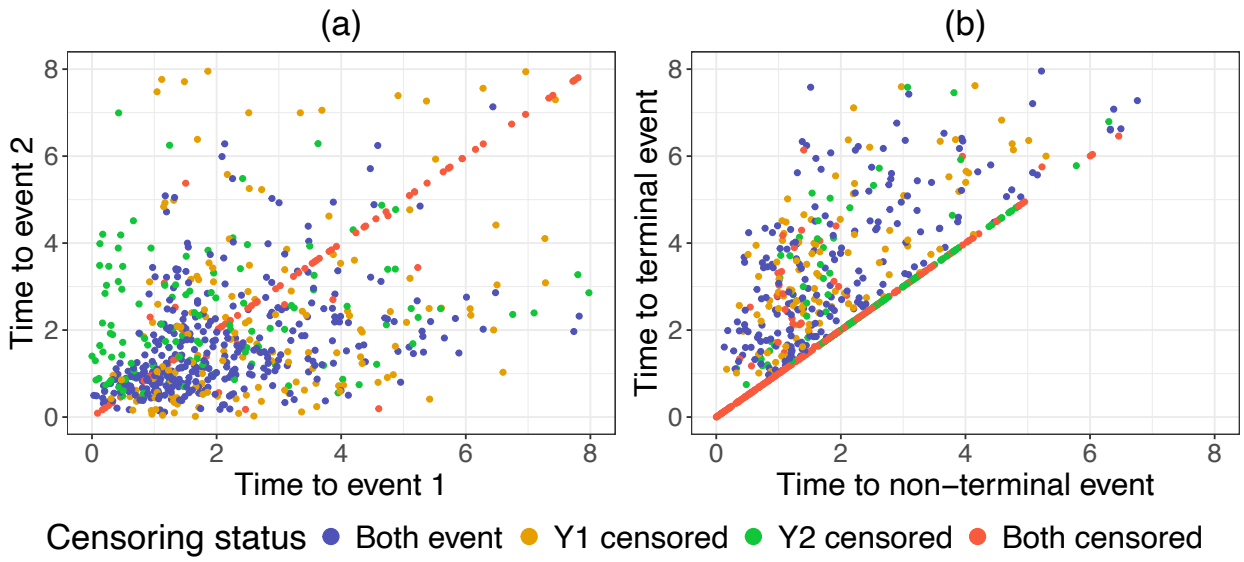

Figure 1: Synthetic bivariate time-to-event data with right-censoring (a) and semi-competing risks (b).

their pair. An example of synthetic bivariate time-to-event data with right-censoring scheme is shown in Figure 1(a). Throughout, we make the common assumption that the marginal censoring times remain independent of their respective true event times as well as from each other.

Moreover, we consider a special type of right-censored time-to-event outcome that naturally produces bivariate data known as "semi-competing risks" (SCR; Fine et al., 2001; Wang, 2003). Semi-competing risks responses usually contain information about a non-terminal and a terminal event. The terminal event may censor the non-terminal one but it remains observable if the non-terminal event occurs first (Fine et al., 2001). In biomedical applications, the terminal event is typically death, whereas the notion of the non-terminal event time is usually a landmark event e.g., time of disease progression. Using our notation, let the true non-terminal and terminal events be denoted by T_1 and T_2, respectively. Semi-competing risks generate bivariate time-to-event data since one observes the first event $Y_1 = \min\{T_1, T_2, T^{cens}\}$ with its corresponding censoring indicator $\delta_1 = \mathbb{1}\{T_1 \leq \min\{T_2, T^{cens}\}\}$. The second observed time-to-event response is then determined by $Y_2 = \min\{T_2, T^{cens}\}$ as well as $\delta_2 = \mathbb{1}\{T_2 \leq T^{cens}\}$ and we again write $(\boldsymbol{Y}, \boldsymbol{\delta})$ for their pair. Figure 1(b) shows a scatterplot of simulated data with semi-competing risks responses.

2.2 Model structure

To describe the entire conditional distribution of right censored time-to-event variables, we make use of a distributional copula regression approach based on generalized additive models for location, scale and shape (GAMLSS; Rigby and Stasinopoulos, 2005). Specifically, we follow Marra and Radice (2020) and Wei et al. (2023) and assume that the joint survival

function $S(t_1, t_2; \boldsymbol{\vartheta}) = P(T_1 > t_1, T_2 > t_2; \boldsymbol{\vartheta})$ is given by

$$S(t_1,\ t_2;\ \boldsymbol{\vartheta}) = C[S_1(t_1; \boldsymbol{\vartheta}^{(1)}),\ S_2(t_2; \boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}], \tag{1}$$

where $C(\cdot, \cdot;\ \vartheta^{(c)}) : [0,1]^2 \to [0,1]$ is a one-parameter bivariate copula function with association parameter $\vartheta^{(c)} \in \mathbb{R}$, and $S_1(t_1; \boldsymbol{\vartheta}^{(1)}) = P(T_1 > t_1; \boldsymbol{\vartheta}^{(1)})$ and $S_2(t_2; \boldsymbol{\vartheta}^{(2)}) = P(T_2 > t_2; \boldsymbol{\vartheta}^{(2)})$ are the possibly different univariate parametric marginal survival functions with respective distribution parameter vectors $\boldsymbol{\vartheta}^{(1)} \in \mathbb{R}^{K_1}, \boldsymbol{\vartheta}^{(2)} \in \mathbb{R}^{K_2}$. Altogether, the bivariate joint survival function depends on the parameter vector $\boldsymbol{\vartheta} = ((\boldsymbol{\vartheta}^{(1)})^\top, (\boldsymbol{\vartheta}^{(2)})^\top, \vartheta^{(c)})^\top \in \mathbb{R}^K$ with $K = K_1 + K_2 + 1$.

Dependence measures An advantage of resorting to copulas is the separation of specifying the marginal distributions and their respective dependence structure. This flexibility could help to uncover important aspects of the association between the marginal event times. In this context, relevant dependence measures are Kendall's τ rank correlation, upper and lower-tail dependence coefficients, and the cross-ratio function. The upper-tail dependence coefficient is defined as $\psi_U = \lim_{q \to 1} P(t_2 > F_2^{-1}(q) | t_1 > F^{-1}(q))$, whereas the lower-tail dependence coefficient is given by $\psi_L = \lim_{q \to 0^+} P(t_2 \leq F_2^{-1}(q) | t_1 \leq F^{-1}(q))$. For instance, the presence of lower-tail dependence would imply that the association between the margins is stronger at the end of the follow-up time (i.e., when $S_1, S_2 \to 0$) and weaker close to the beginning of the study (i.e., when $S_1, S_2 \to 1$), and vice versa for upper-tail dependence. The cross-ratio function is given by

$$R_{\vartheta^{(c)}}(u_1, u_2) = \frac{c(u_1, u_2;\ \vartheta^{(c)})\ C(u_1, u_2;\ \vartheta^{(c)})}{C(u_1 \mid u_2; \vartheta^{(c)})\ C(u_2 \mid u_1; \vartheta^{(c)})}$$

where $c(\cdot, \cdot;\ \vartheta^{(c)})$ denotes the copula density, $u_1 = S_1(t_1; \boldsymbol{\vartheta}^{(1)})$, $u_2 = S_2(t_2; \boldsymbol{\vartheta}^{(2)})$, and the terms $C(u_1 \mid u_2; \vartheta^{(c)}) = \partial C(u_1, u_2;\ \vartheta^{(c)})/\partial u_2$, and $C(u_2 \mid u_1; \vartheta^{(c)}) = \partial C(u_1, u_2;\ \vartheta^{(c)})/\partial u_1$ denote the conditional copula function given the margin u_1 or u_2, respectively. The cross-ratio function provides a measure of local dependence between the margins at S_1, S_2. Values of $R_{\vartheta^{(c)}} > 1$ indicate positive local dependence, whereas $0 < R_{\vartheta^{(c)}} < 1$ points toward negative local dependence. The special case of $R_{\vartheta^{(c)}} = 1$ corresponds to local independence (Emura and Chen, 2018).

Dependence structure We have implemented a wide range of copula functions such as the Gaussian, which is the most prominent example of elliptical copulas, as well as four Archimedean copulas (Frank, Gumbel, Clayton and Joe) with 0, 90°, 180°and 270° rotations of the latter three. Rotating the Clayton, Gumbel and Joe copulas results in changing the direction of the dependence structure to different parts of the quadrant. The three Archimedean copulas and their rotated versions, in contrast to the Gaussian and Frank copulas, do allow for tail dependence.

Marginal survival functions Our implementation features the four most prominent parametric distributions for AFT models: Exponential, Weibull, log-logistic and log-normal. All of the implemented distributions depend on two scalar parameters. Tables A1 and A2 summarise the currently implemented marginal distributions and copula functions, respectively.

2.3 Predictor specifications

Each of the $K = K_1 + K_2 + 1$ parameters of the joint survival function, is modelled as a function of covariates using structured additive predictors $\eta_k^{(\bullet)}$ of the form

$$g_k^{(\bullet)}(\vartheta_k^{(\bullet)}) = \eta_k^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} s_{rk}^{(\bullet)}(\boldsymbol{x}_{rk}), \quad \bullet \in \{1, 2, c\},\ k = 1, \ldots, K_\bullet, \text{ and } K_c = 1, \tag{2}$$

where $\boldsymbol{x}_{rk} \subset \boldsymbol{x}$, and $g_k(\cdot)$ are link functions with corresponding inverse functions $h_k(\cdot) \equiv g_k^{-1}(\cdot)$, guaranteeing that the individual parameters comply with their respective parameter space restrictions. The structured additive predictors $\eta_k^{(\bullet)}$ are composed of a parameter-specific intercept $\beta_{0k}^{(\bullet)}$ and smooth functions of the covariates denoted by $s_{rk}^{(\bullet)}(\cdot)$. The latter can accommodate a wide range of functional forms, such as linear, non-linear and spatial effects. This is because each $s_{rk}^{(\bullet)}(\cdot)$ is modelled through a linear combination of appropriate basis function expansions of the form

$$s_{rk}^{(\bullet)}(\boldsymbol{x}_{rk}) = \sum_{l=1}^{L_{rk}^{(\bullet)}} \beta_{rk,l}^{(\bullet)} B_{rk,l}^{(\bullet)}(\boldsymbol{x}_{rk}),$$

where $B_{rk,l}^{(\bullet)}(\boldsymbol{x}_{rk})$ are the basis functions evaluated at $\boldsymbol{x}_{rk}$ and $\beta_{rk,l}^{(\bullet)}$ are the corresponding unknown regression coefficients which must be estimated, see Wood (2017) for more details.

The summation index $P_k^{(\bullet)}$ in Equation (2) emphasizes that the subset of covariates assigned to each parameter do not need to be the same. In fact, it may be the case that no covariates have an effect on some parameters $\vartheta_k^{(\bullet)}$ of the joint survival function $S(\cdot, \cdot; \boldsymbol{\vartheta})$. Thus, in general there may not be strong a-priori evidence of which subset of covariates (or if any at all) has an effect on the parameters of $S(\cdot, \cdot; \boldsymbol{\vartheta})$. In order to tackle these model-building and variable-selection challenges in a data-driven manner, we resort to component-wise gradient-boosting or statistical boosting to estimate the model coefficients.

2.4 Estimation via component-wise boosting

Statistical boosting (Mayr et al., 2014) is based on a component-wise gradient boosting algorithm with regression-type base-learners (Friedman, 2001; Bühlmann and Hothorn, 2007). In our case, these base-learners correspond to the smooth components $s_{rk}^{(\bullet)}(\boldsymbol{x}_{rk})$, $\bullet \in \{1, 2, c\}$. A complete list of the currently implemented base-learners in the context of boosting can be

found in Mayr et al. (2012). Let $\{(\boldsymbol{y}_i, \boldsymbol{\delta}_i, \boldsymbol{x}_i)\}_{i=1}^n$ be the observed time-to-event data.

Then, estimation of the model coefficients is carried out by iteratively minimizing the empirical risk: $\omega_n = \frac{1}{n}\sum_{i=1}^n \omega(\boldsymbol{y}_i; \boldsymbol{\vartheta}_i)$, where $\boldsymbol{\vartheta}_i = (\boldsymbol{\vartheta}_i^{(1)}, \boldsymbol{\vartheta}_i^{(2)}, \vartheta_i^{(c)}) \in \mathbb{R}^K$ is the distribution parameter vector for observation i, and $\omega(\cdot;\cdot)$ represents the loss function of interest. In our case, the loss is equal to the negative log-likelihood of our model $\mathcal{L} = -\sum_{i=1}^n \ell_i$, where ℓ_i is the log-likelihood contribution. A single contribution to the log-likelihood is given by

$$
\begin{aligned}
\ell = (1-\delta_1)(1-\delta_2)\Big\{ & \log(C[S_1(y_1;\boldsymbol{\vartheta}^{(1)}),\ S_2(y_2;\boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}])\Big\} + \\
(1-\delta_1)\delta_2 & \left\{ \log\left(\frac{\partial C[S_1(y_1;\boldsymbol{\vartheta}^{(1)}),\ S_2(y_2;\boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}]}{\partial S_2(y_2;\boldsymbol{\vartheta}^{(2)})} \right) \ + \ \log(f_2(y_2;\boldsymbol{\vartheta}^{(2)})) \right\} + \\
\delta_1(1-\delta_2) & \left\{ \log\left(\frac{\partial C[S_1(y_1;\boldsymbol{\vartheta}^{(1)}),\ S_2(y_2;\boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}]}{\partial S_1(y_1;\boldsymbol{\vartheta}^{(1)})} \right) \ + \ \log(f_1(y_1;\boldsymbol{\vartheta}^{(1)})) \right\} + \\
\delta_1\delta_2 & \Big\{ \log(c[S_1(y_1;\boldsymbol{\vartheta}^{(1)}),\ S_2(y_2;\boldsymbol{\vartheta}^{(2)});\ \vartheta^{(c)}]) + \log(f_1(y_1;\boldsymbol{\vartheta}^{(1)})) + \log(f_2(y_2;\boldsymbol{\vartheta}^{(2)}))\Big\},
\end{aligned}
\tag{3}
$$

where the functions $f_1(y_1;\boldsymbol{\vartheta}^{(1)})$ and $f_2(y_2;\boldsymbol{\vartheta}^{(2)})$ are the marginal probability density functions (PDFs). In each iteration of the statistical boosting algorithm each of the pre-specified base-learners (components) of each distribution parameters is fitted individually to the negative gradient of the loss function w.r.t. to the additive predictors of the parameters. These quantities are also referred to as pseudo-residuals, and are given by $-\partial\omega(\boldsymbol{y}_i;\boldsymbol{\vartheta}_i)/\partial\eta_{ki}^{(\bullet)}$. Based on a prediction criterion, only the best-performing base-learner or component out of all additive predictors is selected and a "weak" update of the model is conducted (Thomas et al., 2018). The procedure is carried out for a pre-specified number of iterations denoted by $\mathtt{m_{stop}}$. Conducting early stopping, i.e., using $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}} < \mathtt{m_{stop}}$ iterations leads to some base-learners being effectively left out of the model. Hence statistical boosting conducts intrinsic, data-driven variable selection as well as shrinkage of the covariate effects. This implies that the number of fitting iterations $\mathtt{m_{stop}}$ is the main tuning parameter.

Implementation details Our approach extends the boosting methodology presented in Hans et al. (2023) and Briseño Sanchez et al. (2024) to bivariate right-censored time-to-event data. Estimation is carried out in a two-step fashion akin to Joe (2005) described in detail in Algorithm B1. In the first step, the coefficients of the sub-models of the margins are boosted separately, i.e., an optimal number of fitting iterations is obtained for each marginal survival model ($\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(\bullet)}, \bullet = 1, 2$). In the second step, we compute $\hat{S}_\bullet(y_\bullet;\hat{\boldsymbol{\vartheta}}^{(\bullet)})$, as well as $\hat{f}_\bullet(y_\bullet;\hat{\boldsymbol{\vartheta}}^{(\bullet)})$ at the respective $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(\bullet)}$ with $\bullet \in \{1, 2\}$ and plug them into the log-likelihood function shown in Equation (3). The latter is then boosted as a function of $\vartheta^{(c)}$.

For data generated by SCR responses, we proceed similarly but boost only the margin of the terminal event T_2 and compute the fitted survival function and density at $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(2)}$. In the second stage we plug the aforementioned functions into Equation (3) and boost it as a function of $\boldsymbol{\vartheta}^{(1)}$ and $\vartheta^{(c)}$. The algorithm has been integrated into the R package `gamboostLSS`.

We denote our proposed approach described above by `SurvCopBoost`. Section B1 in the Supplementary Material provides an illustration on how to fit the proposed model class using the `SurvCopBoost` function implemented in R.

3 Simulation study

In this section, we conduct a number of experiments to empirically evaluate the estimation accuracy, the predictive performance and the ability of our approach to conduct consistent variable selection. In our experiments, we consider $p_1 = 10$, $p_2 = 500$, $p_3 = 1000$, as well as different censoring regimes in two different scenarios: In Section 3.2 we mimic semi-competing risks (SCR) data with censoring rates similar to those found in our application from Section 4. The simulations in Section 3.3 treat a bivariate time-to-event data (BTE) data generating process (DGP) with "mild" ($\approx 30\%$) and "heavy" ($\approx 70\%$) censoring rates in each margin. Before describing the two scenarios in detail, we state the following general settings that hold for both.

3.1 General settings

Data generation To build the bivariate response distributions we consider the Weibull and log-logistic distributions for the first and second margin, respectively. Bivariate samples from a copula are obtained using the package `VineCopula` Nagler et al. (2022). The copula and predictor choices are scenario-specific and discussed separately. The amount of censoring times also depends on the scenario but in both cases, censoring times are generated independently from univariate distributions. The covariates are generated from a multivariate Gaussian distribution with Toeplitz covariance structure of the form $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p_q$, with $\rho = 0.5$ denoting the correlation between consecutive covariates x_j and x_{j+1}. The range of each covariate is then transformed to the unit interval by means of the standard normal CDF. We generate 500 replicate training data sets of size $n_{\texttt{train}} = 1000$ observations each and evaluate the performance on an additional test set of the same size denoted by $n_{\texttt{test}}$. Since we consider one-parameter copulas and the Weibull and log-logistic distributions come with two distribution parameters each, we have a total of $K = 5$ distribution parameters throughout. Thus, it is worthwhile noting that the cases p_2, p_3 come with 2500 and 5000 potential covariates, such that both can be considered as *high-dimensional* (i.e. $p > n$).

Performance evaluations and benchmarking All performance evaluations are computed using the separate test set. The goodness-of-fit of `SurvCopBoost` is assessed using the negative log-likelihood (log-score) and compared against the respective scores of a competing model assuming the same but independent margins. For further comparison, we evaluate the performance for each margin separately (thus not evaluating the loss in ignoring potential dependence in the responses) in comparison with independent univariate Cox models, as they

represent the most popular approach in survival analysis. To allow for a fair comparison, we used boosting to estimate the Cox models as well. Lastly, we include a penalised maximum likelihood approach implemented in the `GJRM` Marra and Radice (2023) `R` package. The respective criteria are the Integrated Brier Score (IBS), the Integrated Squared Error (ISE), the Integrated Absolute Error (IAE), the Concordance Index (C-Index), as well as true and false positive rates (TPR, FPR, respectively).

Implementation details and tuning To carry out the weak learning mechanism of boosting, we need to set a sensible step-length $\mathtt{s}_{\mathtt{step}}$. Here, we follow Briseño Sanchez et al. (2024) and set $\mathtt{s}_{\mathtt{step}} = 0.1$ for all distribution parameters. However, in order to obtain similar step-lengths among the distribution parameters of the margins, we apply L_2-stabilisation to the parameter-specific gradients (Hofner et al., 2016). We adopt the same step-length for the boosted independent Cox models. The stopping iteration $\mathtt{m}_{\mathtt{stop}}$ of `SurvCopBoost` and the independent Cox models is optimised by minimising the out-of-bag empirical risk on a further validation data set (different from the test data set) of size $n_{\mathtt{mstop}} = 1000$ obtained from the same underlying distribution. We fitted all `SurvCopBoost` models in `R` using our implementations via the `gamboostLSS` package. The boosted Cox models are fitted using the implementation from the package `mboost` (Hothorn et al., 2022). The code to reproduce all results is available on the following GitHub repository: `https://github.com/GuilleBriseno/BoostDistCopReg_Surv`.

3.2 Semi-competing risks (SCR) responses

Data generation Motivated by the data analysed in Section 4, we generate bivariate time-to-event responses that follow the SCR mechanism described in Subsection 2.1 with dependence structure based on a Gumbel copula. Based on the needs of the application, we assume linear predictors given by

$$\begin{aligned}
\log \vartheta_{i1}^{(1)} = \eta_{i1}^{(1)} &= -2x_{1i}, \\
\log \vartheta_{i2}^{(1)} = \eta_{i2}^{(1)} &= +1x_{2i} + 1.5x_{4i}, \\
\log \vartheta_{i1}^{(2)} = \eta_{i1}^{(2)} &= +1x_{1i} + 1.5x_{2i}, \\
\log \vartheta_{i2}^{(2)} = \eta_{i2}^{(2)} &= +1 + 0.75x_{2i} + 0.75x_{4i}, \\
\log(\vartheta_{i}^{(c)} - 1) = \eta_{i}^{(c)} &= 3 - 2x_{2i} - 2x_{4i},
\end{aligned}$$

as well as censoring rates of $\approx 40\%$ and $\approx 47\%$ in each margin, respectively. The censoring times were sampled from a univariate uniform distribution on the interval $[0;\ 7]$. In this case only three out of the p_q, $q \in \{1, 2, 3\}$ covariates have non-zero effects on the distribution parameters. Note that there is an overlap of the informative covariates between the different distribution parameters. The Gumbel copula is able to model upper-tail dependence, hence one would expect larger values of the marginal survival functions (earlier event times) to ex-

hibit a stronger dependence compared to lower values (later event times). Averaging over the observations, the dependence between the margins in terms of Kendall's τ lies within [0.187; 0.922], thus ranging between moderate and very strong positive dependence.

Besides benchmarking with independent models and univariate Cox models, we also compare two ways to estimate `SurvCopBoost`. The first estimates the margins separately using the two-step algorithm described in Algorithm B1 and is denoted as `SurvCopBoost` BTE (*bivariate time-to-event*) estimation. The second estimates first the coefficients that correspond to the margins of the terminal event (T_2). Afterwards, the estimates $\hat{S}_2(\cdot)$ and $\hat{f}_2(\cdot)$ are plugged into Equation (3) and the remainder of the loss is boosted jointly. This procedure is denoted as `SurvCopBoost` SCR (*semi-competing risks*) estimation. We remark that the estimation of the margin corresponding to the terminal event (T_2) is the same for both `SurvCopBoost` BTE and SCR estimation strategies.

Results Table C1 reports the performance metrics. Except the C-Index, all measures are oriented such that lower values indicate better performance. The reported scores are computed as the average of the 500 replicate test data sets. The results emphasize that our proposed `SurvCopBoost` leads to a better fit in terms of the log-score compared to ignoring the dependence structure and fitting independent models. This general observation holds true for both BTE and SCR estimation schemes. However, `SurvCopBoost` SCR appears to outperform the `SurvCopBoost` BTE estimation in terms of the log-score in low-dimensional settings ($p = 10$). In case of high-dimensional data (i.e., $p_2 = 500, p_3 = 1000$), the `SurvCopBoost` BTE strategy outperforms `SurvCopBoost` SCR in terms of the log-score. Univariate performance scores seem to favor `SurvCopBoost` BTE estimation compared to the `SurvCopBoost` SCR approach and also compared to fitting independent Cox models.

Figure C1 displays the estimated linear effects of informative and non-informative covariates in the margin corresponding to the non-terminal event (T_1) as well as the dependence parameter $\vartheta^{(c)}$. In low-dimensional configurations ($p_1 = 10$), both `SurvCopBoost` BTE and `SurvCopBoost` SCR approaches perform similar in $\vartheta_1^{(1)}$ and $\vartheta^{(c)}$. The boxplots in Figure C1, displaying coefficients resulting from `SurvCopBoost` BTE estimation, exhibit a small bias in the intercept as well as the informative covariates in the aforementioned parameters. For $p = 500$ and $p = 1000$ we see that the shrinkage effect on the parameter $\vartheta_2^{(1)}$ becomes stronger the more candidate covariates enter the model. The estimated coefficients of the terminal event are displayed in Figure C2. These boxplots show a similar pattern as those for the non-terminal event, i.e., a stronger shrinkage of the covariate effects on $\vartheta_2^{(2)}$ as p increases.

Regarding the TPRs and FPRs, Table C2 reveals that `SurvCopBoost` BTE estimation tends to select more non-informative covariates in the dependence parameter in low-dimensional configurations than `SurvCopBoost` SCR. On the other hand, for high-dimensional settings

with $p_2 = 500$ or $p_3 = 1000$ potential covariates, `SurvCopBoost` BTE estimation also yields higher TPRs as compared to `SurvCopBoost` SCR. With the most notable differences in the selection rates being observed on the dependence parameter $\vartheta^{(c)}$. The implementation of `GJRM` could only be fitted using $p_1 = 10$ covariates. In that setting the corresponding FPRs were very high due to `GJRM`'s lack of variable selection mechanism. Other results obtained from `GJRM` are omitted.

3.3 Bivariate right-censored time-to-event (BTE) responses

Data generation We consider two censoring regimes with average censoring rates of 30% ("mild") and 70% ("heavy") for both margins, respectively. The bivariate observations are generated from a Clayton copula, which allows to model positive dependence as well as lower tail dependence between the margins. We consider two DGPs. The first DGP contains only linear effects of the covariates, whereas the second DGP consists of non-linear effects. For these, the additive predictors are

Linear DGP:	Non-linear DGP:
$\log \vartheta_{1i}^{(1)} = \beta_{0,1}^{(1)} - 2x_{1i},$	$\log \vartheta_{1i}^{(1)} = -1.8\cos(4x_{3i}),$
$\log \vartheta_{2i}^{(1)} = +1x_{2i} + 1.5x_{4i},$	$\log \vartheta_{2i}^{(1)} = 0.02 - \sin(x_{1i}) + \exp(x_{1i}+1)^2 + 3\cos(2\pi x_{1i}),$
$\log \vartheta_{1i}^{(2)} = \beta_{0,1}^{(2)} + 1x_{1i} + 1.5x_{2i},$	$\log \vartheta_{1i}^{(2)} = 2\sin(4x_{2i}),$
$\log \vartheta_{2i}^{(2)} = \beta_{0,2}^{(2)} + 0.75x_{2i} + 0.75x_{4i},$	$\log \vartheta_{2i}^{(2)} = -0.979\cos(2x_{4i}) - 1.958\tanh(x_{4i}),$
$\log \vartheta_{i}^{(c)} = 3 - 2x_{2i} - 2x_{4i},$	$\log \vartheta_{i}^{(c)} = -3.1\cos(4x_{3i}).$

Consequently, only three/four out of the p_q, $q \in \{1, 2, 3\}$ covariates have non-zero effects on the distribution parameters in the linear/non-linear DGPs, respectively. Furthermore, several of the few informative covariates have an effect on multiple distribution parameters which challenges estimation. For the linear DGP, the additive predictor of the dependence parameter $\vartheta^{(c)}$ covers Kendall's τ values within $[0.159;\ 0.907]$, whereas for the non-linear DGP it ranges from $[0.022;\ 0.917]$. Thus covering from low to very strong positive dependence between T_1 and T_2 in both DGPs. In addition, the chosen intercepts $\beta_{0,1}^{(2)}$ paired with independent censoring times sampled from uniform distributions on $[0;\ 8.5]$ yield censoring rates of about 30% and 70% for the linear DGP. In the non-linear DGP, the mild censoring regime is obtained by using uniform distributions on $[0; 11]$, whereas the heavy censoring regime uses the interval $[0; 2.75]$ for sampling the censoring times.

Results for the linear DGP Table C3 reports the log-scores. The difference in log-scores between `SurvCopBoost` and independent models starts to dissipate only in extreme cases with a very high number of potential covariates ($p_3 = 1000$) and heavy censoring in the margins (70%), see column (2), $p_3 = 1000$. In line with these findings, `SurvCopBoost` also produces better univariate scores compared to the univariate Cox models. The estimated coefficients are shown in Figure C3. Given a mild censoring rate (30% in each margin) and a low number of potential covariates ($p_1 = 10$), `SurvCopBoost` recovers the effect of informative covariates

quite well, although the shrinkage of effect estimates is stronger for the dependence parameter $\vartheta^{(c)}$. On one hand, increasing the number of potential covariates as well as increasing the censoring rate (70%) has a negligible effect on the estimation of informative covariates on the distribution parameters $\vartheta_1^{(1)}, \vartheta_1^{(2)}$. On the other hand, the shrinkage of effect estimates increases sharply in the parameter $\vartheta_2^{(2)}$ as well as the dependence parameter $\vartheta^{(c)}$. The parameter $\vartheta_1^{(2)}$ also exhibits considerable shrinkage of the effect estimates on high-dimensional settings and heavy censoring, although it is not as pronounced as on the two aforementioned parameters.

The TPRs and FPRs presented in the upper half of Table C4 show that `SurvCopBoost` is able to accurately recover the effect of informative covariates across the studied configurations. It can be seen that the degree of shrinkage and regularization depends more on the censoring rate than on the number of potential covariates present in the data, e.g., compare the TPR in columns (1) against (2) for $p_3 = 1000$ in Table C4. Similar to Section 3.2, the implementation of `GJRM` could only be fitted in configurations with $p_1 = 10$ covariates. The respective FPRs exhibited the same pattern as in Section 3.2. Once again, further results obtained using `GJRM` are omitted.

Results for the non-linear DGP Similar to the linear DGP, `SurvCopBoost` outperforms the independent models in terms of the log-score in almost all considered configurations. Under a high censoring rate (70%) combined with a high number of potential covariates ($p_2 = 500, p_3 = 1000$ in Table C3) the performance of both models is similar. This behaviour can be also observed in some of the univariate scores such as the IBS and C-Index, where those produced by Cox models are slightly better than those from `SurvCopBoost`. The estimated non-linear effects of the informative covariates shown in Figure C4 indicate that the censoring rates and the increasing number of potential covariates have a negligible effect on the accuracy of the estimated effects on the parameters of the marginal survival functions. However, increasing amount of censoring and noise variables induces a stronger shrinkage of the estimated effects and thus a larger bias in the dependence parameter $\vartheta^{(c)}$. For example, the green curves in Figure C4 corresponding to the row showing 70% censoring exhibit a flatter shape of the estimated non-linear effect compared to the row depicting 30% censoring.

The selection rates corresponding to the non-linear DGP are shown in the lower half of Table C4. As already established in the linear DGP, `SurvCopBoost` identifies the informative covariates in all parameters of the joint survival function regardless of the number of candidate covariates in the model in a mild censoring regime (30% censoring). The FPR in low-dimensional settings are rather high for both `SurvCopBoost` and independent Cox models, but they rapidly shrink towards zero once a large number of candidate covariates enter the model. However, the TPR from the Cox models is considerably lower than those of `SurvCopBoost`. Results from `GJRM` are once again omitted and the FPR behave in the same

way as described in the results of Section 3.2.

3.4 Summary of the simulation results

Overall, `SurvCopBoost` demonstrated satisfactory results for both SCR and BTE data. It is able to effectively detect and recover all true effects across the distribution parameters of the bivariate distribution. However, a larger bias in the estimation of the dependence parameter under heavy-censoring has to be acknowledged. This is likely because the copula dependence parameter $\vartheta^{(c)}$ shows stronger shrinkage of informative effects compared to other parameters. The strength of induced shrinkage and regularization is also influenced by the censoring rate and the number of candidate variables. This phenomenon may be attributed to the greedy nature of the algorithm, since a reduction of the loss from including a covariate with a small coefficient in the dependence parameter might not be large enough compared to updating a coefficient in any other parameter corresponding to the margins or even the intercept of $\vartheta^{(c)}$, i.e. constant dependence.

In high-dimensional SCR configurations, such as the one analysed in Section 4, our proposed two-step estimation approach (`SurvCopBoost` BTE) performs well at identifying informative covariates as well as modeling the underlying bivariate distribution. Overall, evaluating the predictive behaviour via probabilistic scores highlights the added value of the bivariate `SurvCopBoost` model compared to using boosting for independent AFT models or more traditional Cox models for bivariate time-to-event data. Compared to the penalised maximum likelihood approach of `GJRM`, the proposed `SurvCopBoost` allows not only for a more streamlined model-building process by selecting the most informative variables in a data-driven manner, but also for feasible estimation in high-dimensional ($p \gg n$) settings.

4 Analysis of high-dimensional ovarian cancer data with semi-competing risks responses

In this section we showcase the ability of the proposed `SurvCopBoost` to conduct data-driven variable selection in a challenging high-dimensional data structure with semi-competing risks responses. The data analysis is related to ovarian cancer, a leading cause of cancer death in women (Siegel et al., 2020) and the second global cause of death from gynecologic cancers (Bai et al., 2020). We are concerned with estimating the joint survival function of the time to tumour progression, i.e., a landmark event of the disease, and the time of death. Using `SurvCopBoost`, the parameters of the joint survival function are modelled as functions of informative covariates selected in a data-driven fashion from a high-dimensional covariate vector. The data were obtained from the `R` Bioconductor package `curatedOvarianData` (Ganzfried et al., 2013). Next, we describe the data extraction process, configurations used

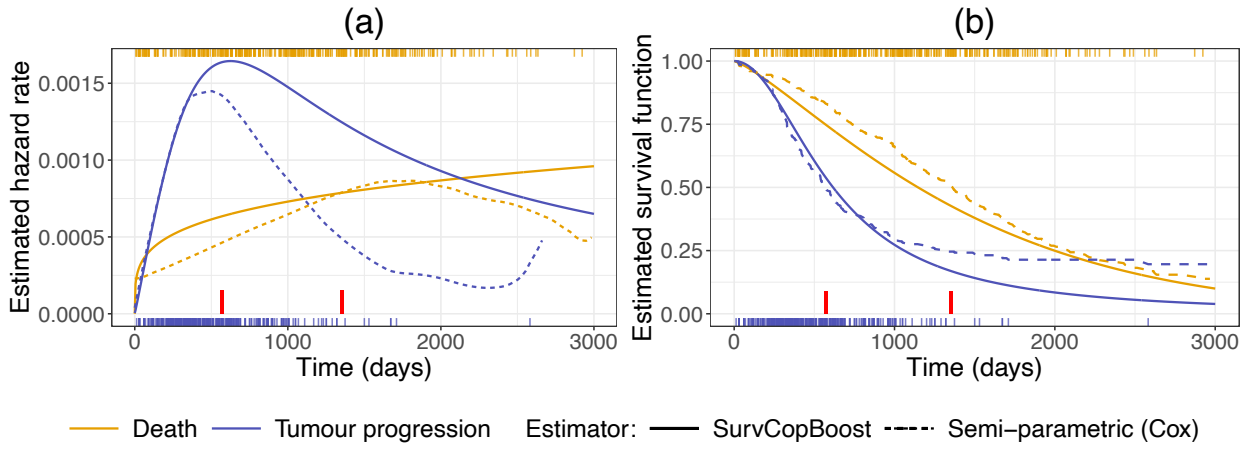

Figure 2: Estimated baseline hazard rate (a) and survival function (b) for time of tumour progression and time of death. Solid lines are estimates from `SurvCopBoost`, whereas dashed lines denote semi-parametric estimates corresponding to independent univariate Cox models. Thick red vertical lines highlight the median time of tumour progression (570 days) and median time of death (1353 days).

for the `SurvCopBoost` model, as well as the results of our analysis.

Data structure The data is comprised of four annotated studies (GSE17260, GSE30161, GSE9891, and TCGA) included in the `curatedOvarianData` (Ganzfried et al., 2013) package. The studies were extracted according to the `patientselection.config` file, see the package's vignette for more details. Our extracted sample consists of a total of $n = 822$ patients. Following a semi-competing risks data generating process, the responses are given by each patient's time of tumour progression (non-terminal event, T_1) and their respective time of death or survival time (terminal event, T_2) after surgery. The time scale of the responses is given in days. The median time-to-event times are 570 and 1353 days, respectively. The censoring rate for tumour progression is $\approx 40\%$, whereas in $\approx 48\%$ of patients the terminal event was not observed. These censoring rates are similar to those considered in our simulations under an SCR DGP. We consider all the covariate information that is commonly available across the aforementioned studies. Information from the common covariates may be split into two types: genomic and clinical. The regressors containing genomic information are a total of 11,761 gene expressions. Following Ganzfried et al. (2013) as well as Emura et al. (2018), the independent variables with clinical information are the tumour stage according to the FIGO staging system (I-IV, dummy encoding) and the residual tumour size at surgery encoded as a dummy variable as well (0= under 1cm, 1 = over 1cm). This yields a total of $p = 11{,}763$ covariates, which corresponds to a high-dimensional setting. In fact, fitting a statistical model to such a data structure ($p \gg n$) is infeasible with standard techniques. A previous analysis conducted on similar data by Emura et al. (2018) carried out variable selection based on univariate hypothesis testing prior to model fitting (Jenssen et al., 2002). Their approach selected 158 gene expressions associated with the non-terminal event (T_1), and 128 genes for time of death (T_2) out of the same set of potential covariates we examine here. Afterwards a

composite covariate (Tukey, 1993) is taken as a summary of the selected "most significant" variables. In our case `SurvCopBoost` allows the entire covariate vector to enter the model directly.

Model configuration and tuning We split our extracted sample into three partitions. The training data ($n_{\texttt{train}} = 577$), the validation data for tuning the number of fitting iterations ($n_{\texttt{mstop}} = 128$) and the data determining the optimal marginal distributions and copula function by means of the out-of-sample log-score ($n_{\texttt{test}} = 117$), respectively. The log-logistic, log-normal and Weibull distributions are considered as candidates for the margins of each of the event-times. For the dependence structure we fit 14 different implemented copula functions to the best-fitting marginal distributions. The copulas are the Gaussian, Frank, Clayton, Gumbel and Joe, as well as 90°, 180°and 270° rotations of the latter three. In total, the joint survival function consists of five parameters. The following additive predictor configuration is used for all parameters of the joint survival function:

$$\eta_k^{(\bullet)} = \beta_{0k}^{(\bullet)} + \sum_{r=1}^{P_k^{(\bullet)}} x_{rk}\beta_{rk}^{(\bullet)}, \quad \bullet \in \{1, 2, c\},\ k = 1, \ldots, K_\bullet, \text{ and } K_c = 1, \tag{4}$$

where x_{rk} denotes one of the $p = 11{,}763$ covariates in the data. Hence, all covariates are modelled as linear functions. To the best of our knowledge, this is the first instance where the entire covariate vector is considered for modelling of this data. We determine the best-fitting marginal distributions and copula function by means of the out-of-sample log-score.

Due to the relatively small sample size used for estimating the model coefficients ($n_{\texttt{train}} = 577$), we set the step-length to $\texttt{s}_{\texttt{step}} = 0.005$. This configuration will lead to a larger number of optimal iterations, but it will keep the boosting algorithm stable throughout the fitting process. We apply L_2 stabilisation to the negative gradients of the loss and fit `SurvCopBoost` as stated before. Lastly, we fit independent univariate Cox models using boosting to each of the time-to-event responses for comparison.

Results The best-fitting distribution for time to tumour progression is the log-logistic distribution, whereas for time of death it is the Weibull distribution. This result points to the difference in statistical behaviour between the time of tumour progression and the survival time. Figure 2(a) shows the estimated baseline hazard rates as well as baseline survival functions in (b). An important aspect is the mode of the hazard of time to tumour progression which can be seen to occur within the first 1000 days. This indicates a higher risk of tumour progression earlier after surgery compared to later in time. In contrast, the estimated baseline hazard of time to death has a monotonic increasing shape. The estimated baseline survival functions reveal the lower median time-to-event for the non-terminal event compared to death. Thus the drop in progression-free survival is much sharper compared to the terminal

Table 1: Out-of-sample log-scores of candidate marginal distributions and copula functions. Best-fitting values highlighted with bold numbers.

Selection of marginal distributions

Distribution	tumour progression (Non-terminal event, T_1)	Death (Terminal event, T_2)
Weibull	571.48	**485.73**
Log-logistic	**550.45**	485.98
Log-normal	555.69	506.22

Selection of copula function

	Copula	log-score		Copula	log-score
1	Independence	1036.18	9	Clayton 90°	1037.78
2	Gaussian	1016.71	10	Gumbel 90°	1037.42
3	Clayton	1022.92	11	Joe 90°	1037.16
4	Clayton 180°	1017.07	12	Clayton 270°	1037.33
5	Gumbel	**1014.70**	13	Gumbel 270°	1037.78
6	Gumbel 180°	1019.01	14	Joe 270°	1037.82
7	Joe	1017.39	15	Frank	1020.41
8	Joe 180°	1023.53			

log-scores computed using $n_{\texttt{test}} = 117$ observations.

event. The estimated semi-parametric baseline hazard and functions that correspond to the Cox model follow those estimated by `SurvCopBoost` when there is a high prevalence of observations. The semi-parametric estimators show lower hazards in regions without observations, however this behaviour is expected in estimators of this type, see the rugs in Figure 2(a) and (b). A similar phenomenon can be seen in the estimated semi-parametric baseline survival functions (dashed lines) in Figure 2(b).

A total of 95 covariates for the model of time to tumour progression (non-terminal event) is selected, see Table 2. More specifically, it selects 73 variables for the parameter $\vartheta_1^{(1)}$ and 24 variables for $\vartheta_2^{(1)}$ with only two genomic variables overlapping. The binary variable `residual tumour size` was the only clinical covariate selected for the sub-model $\vartheta_1^{(1)}$ of tumour progression. Our proposed `SurvCopBoost` and the significance-testing-based variable selection approach from Emura et al. (2018) have an overlap of 22 gene expressions. Out of these 22 overlapped variables, `SurvCopBoost` selects six of the top ten "most significant" expressions.

Previous analyses and meta-analyses have shown the expression of gene $CXCL12$ (encoding a chemokine related to immune response) to be associated with survival (Popple et al., 2012; Ganzfried et al., 2013; Emura et al., 2017). Albeit these studies focused exclusively on this particular gene while ignoring others. In this case `SurvCopBoost` selected $CXCL12$ only for

Table 2: Number of selected covariates and optimal fitting iterations of the parameters of the joint survival function using `SurvCopBoost` as well as boosted univariate independent Cox models. The symbols λ_1 and λ_2 denote the hazard rate corresponding to each Cox model.

	Time of tumour progression (T_1) Log-logistic distribution		Time of death (T_2) Weibull distribution		Dependence Gumbel copula	Cox T_1	Cox T_2
	$\vartheta_1^{(1)}$	$\vartheta_2^{(1)}$	$\vartheta_1^{(2)}$	$\vartheta_2^{(2)}$	$\vartheta^{(c)}$	λ_1	λ_2
Selected covariates	73	24	26	8	1	69	115
$\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}}$	1740	2165	824	1313	18	2594	7487
Link	$\ln(\cdot)$	$\ln(\cdot)$	$\ln(\cdot)$	$\ln(\cdot)$	$\ln(\cdot - 1)$	$\ln(\cdot)$	$\ln(\cdot)$

L_2 stabilisation, $\mathtt{s}_{\mathtt{step}} = 0.005$, $n_{\mathtt{train}} = 577$, and $n_{\mathtt{mstop}} = 128$.

the parameter $\vartheta_1^{(1)}$ of time to tumour progression's distribution. The association of this gene expression with the non-terminal event is also confirmed by Emura et al. (2018). Other selected genes include members of the $TIMP$ family ($TIMP2$), which have functions associated with cell proliferation and survival Bourboulia et al. (2011). The expression $PTPN4$, which has been found to perform an essential role in most phenotypes of tumour cells (Tang et al., 2022), was selected in both parameters $\vartheta_1^{(1)}$ and $\vartheta_2^{(1)}$ of the non-terminal event's distribution. Another gene selected in the aforementioned parameters was $FAT2$, which according to Wang et al. (2022) shows promise to be a predictor for responsiveness to immunotherapy and prognosis in uterine corpora malignant tumours. The gene $HIST1H4E$, selected for $\vartheta_2^{(1)}$, has been found to play a role in the production of $CD8^+$ regulatory T-cells or pathogen-combating cells (Wu et al., 2016).

For the distribution of the survival time a total of 34 covariates were selected. As shown in Table 2, out of the informative variables for the terminal event, 26 were selected for $\vartheta_1^{(2)}$ and eight for $\vartheta_2^{(2)}$, respectively. In this case there was no overlap in the selected covariates across the parameters. As previously mentioned, the univariate significance-testing-based variable selection approach used in Emura et al. (2018) identified a total of 128 genes with time of death, which is a slightly sparser model compared to that of the non-terminal event (tumour progression). In our case we observe a similar pattern of a sparser model for the time of death. `SurvCopBoost` has twelve gene expressions in common with the approach from Emura et al. (2018) and features once again six variables of the top ten "most significant" ones. An important expression that was selected out of the most significant ones from Emura et al. (2018) is $TEAD1$. It has been found that the $TEAD$ genetic family is abnormally expressed in patients with Ovarian Serous Carcinoma (Ren et al., 2021), which is the most common type of ovarian cancer (Ovarian Cancer Research Alliance , OCRA). The selected expression of gene $YWHAB$ is associated with advanced stages of ovarian cancer as well as poor patient prognosis Li et al. (2021). Our proposed `SurvCopBoost` selects $VSIG4$ into the sub-model $\vartheta_2^{(2)}$. It has been found that $VSIG4$ shows over-expression in ovarian cancers compared with benign tumours and could be a potential target for therapy Byun et al. (2017).

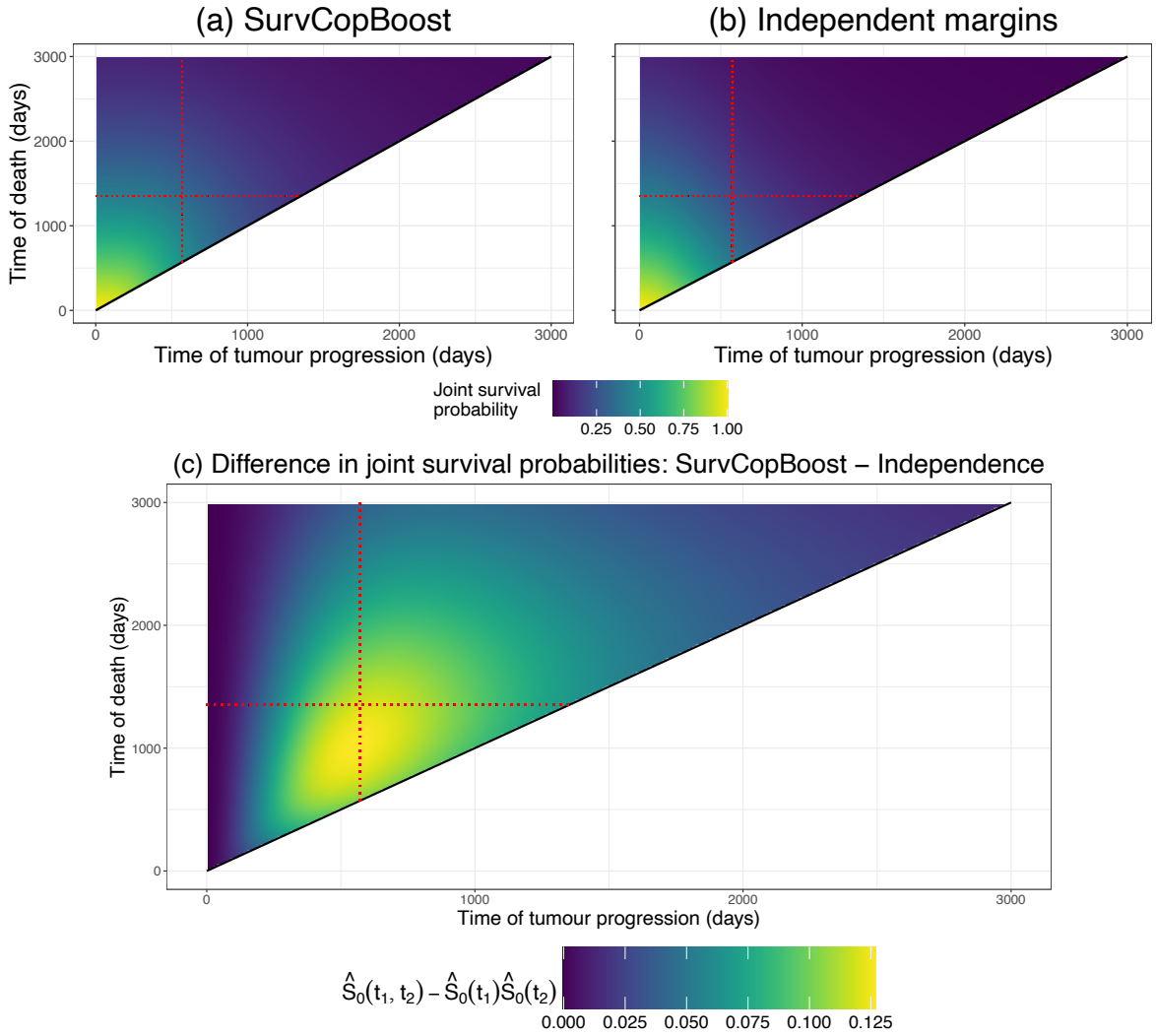

Figure 3: Estimated baseline joint survival probability of time of tumour progression and time of death in days with Gumbel copula using `SurvCopBoost` (a) as well as independent Log-logistic and Weibull margins (b). Difference between the baseline joint survival functions obtained using `SurvCopBoost` and independent margins, i.e. $\hat{S}_0(t_1, t_2; \hat{\boldsymbol{\vartheta}}) - \hat{S}_0(t_1; \hat{\boldsymbol{\vartheta}}^{(1)})\hat{S}_0(t_2; \hat{\boldsymbol{\vartheta}}^{(2)})$ (c). Red dotted lines indicate the median event-times: 570 days for tumour progression and 1353 days for death, respectively. Only the upper-wedge is defined for SCR data.

The Gumbel copula is selected as best-fitting dependence function, see Table 1. This table furthermore reveals that the data strongly rejects copulas that support dependence for large values of time such as the Clayton, Gumbel 180°, or Joe 180°. Copulas that support negative dependence are strongly rejected as well. This can be seen in the worse predictive performance compared to that of a model with independent margins, see the log-score corresponding to 90° and 270° rotations. Figure 3(a) depicts the estimated baseline joint survival function according to the Gumbel copula model with log-logistic distributed time to tumour progression and Weibull distributed time of death. It can be seen that the joint survival is rather high for the first 100 days after surgery. A decrease in joint survival can be seen after 1000 days. The joint survival function assuming independent margins is shown in Figure 3(b). It can be seen that for regions close to the median event times the joint survival function assuming independence

exhibits lower joint survival probabilities, compared to that of `SurvCopBoost`. The difference between the estimated joint survival functions, i.e. $\hat{S}_0(t_1, t_2; \hat{\boldsymbol{\vartheta}}) - \hat{S}_0(t_1; \hat{\boldsymbol{\vartheta}}^{(1)})\hat{S}_0(t_2; \hat{\boldsymbol{\vartheta}}^{(2)})$, is depicted in Figure 3(c). This shows that the joint survival probability of tumour progression and death is underestimated when both event times are modelled independently, with the biggest discrepancy between the estimates being observed close to the median event times, see the bright yellow spot around the intersection of the red dotted lines in Figure 3(c).

Only one gene expression ($SLC16A10$) is selected for the model of $\vartheta^{(c)}$. This covariate was neither selected in the model of time to tumour progression nor in the one of time to death. Members of the $SLC16A$ gene family are important for cell metabolism (Halestrap and Meredith, 2004) and are known to play a crucial role in the process of tumourigenesis, i.e., the formation of cancer as well as tumour progression (Yu et al., 2020). This particular variant was not selected by the significance-testing-based heuristic employed in Emura et al. (2018). `SurvCopBoost` allows us to compute dependence measures in order to gain additional insights of the relationship between the margins. The estimated baseline dependence between the margins expressed as Kendall's τ is $\hat{\tau} = 0.5$ and taking $SLC16A$ into consideration yields values of $\hat{\tau} \in [0.496;\ 0.510]$, indicating a moderate dependence between time to tumour progression and survival time. This result aligns with the estimated dependence previously found by Emura et al. (2017) and Emura et al. (2018). Additionally, the Gumbel copula supports upper-tail dependence, thus meaning that the margins are dependent for extremely high values of their respective survival functions, i.e., at very early times. This result is clinically reasonable, since patients that unfortunately suffer from tumour progression early after surgery typically also have a poorer prognosis of overall survival.

The range of the estimated upper-tail dependence coefficients in the data is $\hat{\psi}_U \in [0.582;\ 0.595]$. This shows that the margins are moderately dependent at extremely early times. In fact, the upper-tail dependence is higher than the dependence quantified by the estimated Kendall's τ. Lastly, the values of the estimated cross-ratio function $\hat{R}_{\vartheta^{(c)}}$ show that the local dependence between the margins is always positive and becomes very high for some observations. The range of the estimated function is within $\hat{R}_{\vartheta^{(c)}} \in [1.230;\ 540.092]$ and has a median of $\text{med}(\hat{R}_{\vartheta^{(c)}}) = 2.333$.

5 Discussion

We have introduced `SurvCopBoost`, which is a distributional copula regression approach for bivariate time-to-event data under right-censoring and for semi-competing risks. Estimation in `SurvCopBoost` is carried out via statistical boosting (Bühlmann and Hothorn, 2007). This enables data-driven variable selection, a feature that considerably simplifies the complex model building process. Our simulation studies show that `SurvCopBoost` outperforms

other approaches (independent univariate boosted Cox and AFT, as well as bivariate copula time-to-event using penalised maximum likelihood) in terms of probabilistic forecast and exhibits similar performance to its competitors in terms of univariate metrics. `SurvCopBoost` also performs satisfactory in terms of variable selection by being able to identify informative covariates, as reflected TPRs and FPRs. All of these qualities were observed under different censoring regimes and growing number of noise variables in the model.

We analysed a high-dimensional data structure extracted from the R Bioconductor package `curatedOvarianData` (Ganzfried et al., 2013) with time-to-event responses following a semi-competing risks data generating process. `SurvCopBoost` selected a subset of 129 informative covariates for the distributions of the marginal event times out of a potential 11,763 variables. Therefore, `SurvCopBoost` demonstrates the benefit of conducting data-driven variable selection by analysing jointly the *entire* covariate vector instead of relying on heuristics, for example hypothesis testing performed on univariate regression models. We believe that our application presented in Section 4 demonstrates the advantages of using `SurvCopBoost` for analysing challenging data structures in a time-to-event analysis context.

Currently `SurvCopBoost` implements three parametric distributions: Weibull, log-logistic and log-normal. The implementation of "umbrella" distributions, such as the generalised gamma (Cox et al., 2007) or generalised F distributions (Cox, 2008), which contain the already implemented ones as special cases, could be an option to further extend the flexibility of `SurvCopBoost`. A potential caveat of the current implementation of `SurvCopBoost` is the distributional assumption of a specific family for the marginal event times. Identifying a suitable distribution might be challenging in some cases. A pragmatic solution could be to implement Cox-type margins (Deresa and Keilegom, 2024) or fully non-parametric margins (Akritas, 2004). However, we consider link-based or "generalised time-to-event models" (Liu et al., 2018; Marra and Radice, 2020) to be a more appropriate approach since those models are based on semi-parametric regression techniques.

We are currently exploring the inclusion of cure fractions, i.e., cure models (Othus et al., 2012; Peng and Yu, 2021), to account for observations that do not experience the event of interest, or in other words, their survival function does not reach zero. For example, this can be the case in semi-competing risk data where there are individuals that will not experience the landmark or non-terminal event. Combining statistical boosting and cure models can be very beneficial, since it is likely that some covariates will have an effect on the cure fraction and not on the survival function or vice versa. Therefore a purely data-driven variable selection mechanism could simplify the model building process. Other areas of active research are the censoring scheme and mechanism or their underlying assumptions thereof. We are interested in adapting a more general censoring scheme, which would allow to model data that features not only right, but also left and interval-censored observations, see e.g., Sun and Ding (2019)

or Petti et al. (2022). Regarding the censoring mechanism, the validity of the independent, as well as non-informative censoring in the marginal responses can be put up to debate / openly challenged or questioned. Allowing for dependent censoring in the marginal responses would require us to model the dependence structure between the marginal censoring and event times, see e.g., Czado and Van Keilegom (2022). Informative censoring could be addressed by adapting the approach of Dettoni et al. (2020) to the framework of `SurvCopBoost`. These developments would result in a more complex model structure but will ultimately be beneficial for practical data analysis.

The boosting algorithm underlying `SurvCopBoost` is prone to some shortcomings. One of these aspects is the rather high FPRs, i.e. including non-informative explanatory variables in the model, in particular in low-dimensional settings. De-selection of non-informative covariates as proposed by Strömer et al. (2022) for statistical boosting could be adopted in `SurvCopBoost`. The use of a constant step-length throughout the fitting process in gradient boosting can lead to a slow convergence of the algorithm as pointed out by Zhang et al. (2022). Since the joint survival functions set up by `SurvCopBoost` feature a large number of distribution parameters, an adaptive step-length as proposed by Zhang et al. (2022) or Daub et al. (2024) would lead to considerable improvements in this area.

Acknowledgements

The work on this article was supported by the German research foundation (DFG) through the grants KL3037/2-1, MA7304/1-1 (428239776).

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

References

Akritas, M. G. (2004). Nonparametric survival analysis. *Statistical Science*, 19(4):615–623.
URL: `http://www.jstor.org/stable/4144432`.

Bai, J., Xie, Z., and Sun, L. (2020). Case report: Metachronous quadruple cancers including breast cancer and triple genital cancer. *International Journal of General Medicine*, 13:1575–1580.
URL: `http://dx.doi.org/10.2147/IJGM.S278219`.

Beis, G., Iliopoulos, A., and Papasotiriou, I. (2024). An overview of introductory and advanced survival analysis methods in clinical applications: Where have we come so far? *Anticancer Research*, 44(2):471–487.
URL: https://www.doi.org/10.21873/anticanres.16835.

Beyersmann, J., Melis, G. G., Kneib, T., Molenberghs, G., Muggeo, V., Vansteelandt, S., and Heller, G. Z. (2024). Discussion on: 'simple or complex statistical models: non-traditional regression models with intuitive interpretations' by Gillian Z. Heller. *Statistical Modelling*, page 1471082X241277642.
URL: https://doi.org/10.1177/1471082X241277642.

Binder, H., Allignol, A., Schumacher, M., and Beyersmann, J. (2009). Boosting for high-dimensional time-to-event data with competing risks. *Bioinformatics*, 25(7):890–896.
URL: https://www.doi.org/10.1093/bioinformatics/btp088.

Bourboulia, D., Jensen-Taubman, S., Rittler, M. R., Han, H. Y., Chatterjee, T., Wei, B., and Stetler-Stevenson, W. G. (2011). Endogenous angiogenesis inhibitor blocks tumor growth via direct and indirect effects on tumor microenvironment. *The American Journal of Pathology*, 179(5):2589–2600.
URL: https://doi.org/10.1016/j.ajpath.2011.07.035.

Briseño Sanchez, G., Klein, N., Klinkhammer, H., and Mayr, A. (2024). Boosting distributional copula regression for bivariate binary, discrete and mixed responses.
URL: https://arxiv.org/abs/2403.02194.

Bühlmann, P. and Hothorn, T. (2007). Boosting algorithms: Regularization, prediction and model fitting. *Statistical Science*, 22(4):477–505.
URL: https://doi.org/10.1214/07-STS242.

Byun, J. M., Jeong, D. H., Choi, I. H., Lee, D. S., Kang, M. S., Jung, K. O., Jeon, Y. K., Kim, Y. N., Jung, E. J., Lee, K. B., Sung, M. S., and Kim, K. T. (2017). The significance of VSIG4 expression in ovarian cancer. *International Journal of Gynecologic Cancer*, 27(5):872–878.
URL: https://doi.org/10.1097/IGC.0000000000000979.

Chowdhury, M. Z. I. and Turin, T. C. (2020). Variable selection strategies and its importance in clinical prediction modelling. *Family Medicine and Community Health*, 8(1):e000262.
URL: https://www.doi.org/10.1136/fmch-2019-000262.

Cox, C. (2008). The generalized F distribution: An umbrella for parametric survival analysis. *Statistics in Medicine*, 27(21):4301–4312.
URL: https://doi.org/10.1002/sim.3292.

Cox, C., Chu, H., Schneider, M. F., and Muñoz, A. (2007). Parametric survival analysis and taxonomy of hazard functions for the generalized gamma distribution. *Statistics in*

Medicine, 26(23):4352–4374.
URL: https://doi.org/10.1002/sim.2836.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 34(2):187–202.
URL: https://doi.org/10.1111/j.2517-6161.1972.tb00899.x.

Czado, C. and Van Keilegom, I. (2022). Dependent censoring based on parametric copulas. *Biometrika*, 110(3):721–738.
URL: https://doi.org/10.1093/biomet/asac067.

Daub, A., Mayr, A., Zhang, B., and Bergherr, E. (2024). A balanced statistical boosting approach for GAMLSS via new step lengths.
URL: https://arxiv.org/abs/2404.08331.

De Bin, R. and Stikbakke, V. G. (2023). A boosting first-hitting-time model for survival analysis in high-dimensional settings. *Lifetime Data Analysis*, 29(2):420–440.
URL: https://www.doi.org/10.1007/s10985-022-09553-9.

Deresa, N. W. and Keilegom, I. V. (2024). Copula based Cox proportional hazards models for dependent censoring. *Journal of the American Statistical Association*, 119(546):1044–1054.
URL: https://doi.org/10.1080/01621459.2022.2161387.

Dettoni, R., Marra, G., and Radice, R. (2020). Generalized link-based additive survival models with informative censoring. *Journal of Computational and Graphical Statistics*, 29(3):503–512.
URL: https://doi.org/10.1080/10618600.2020.1724544.

Emura, T. and Chen, Y.-H. (2018). *Analysis of Survival Data with Dependent Censoring: Copula-Based Approaches*. Springer Singapore.
URL: https://www.doi.org/10.1007/978-981-10-7164-5.

Emura, T., Nakatochi, M., Matsui, S., Michimae, H., and Rondeau, V. (2018). Personalized dynamic prediction of death according to tumour progression and high-dimensional genetic factors: Meta-analysis with a joint model. *Statistical Methods in Medical Research*, 27(9):2842–2858.
URL: https://doi.org/10.1177/0962280216688032.

Emura, T., Nakatochi, M., Murotani, K., and Rondeau, V. (2017). A joint frailty-copula model between tumour progression and death for meta-analysis. *Statistical Methods in Medical Research*, 26(6):2649–2666.
URL: https://doi.org/10.1177/0962280215604510.

Fine, J. P., Jiang, H., and Chappell, R. (2001). On semi-competing risks data. *Biometrika*, 88(4):907–919.
URL: http://www.jstor.org/stable/2673691.

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232.

Ganzfried, B. F., Riester, M., Haibe-Kains, B., Risch, T., Tyekucheva, S., Jazic, I., Wang, X. V., Ahmadifar, M., Birrer, M. J., Parmigiani, G., Huttenhower, C., and Waldron, L. (2013). curatedOvarianData: Clinically annotated data for the ovarian cancer transcriptome. *Database*, 2013.
URL: http://dx.doi.org/10.1093/database/bat013.

Griesbach, C., Groll, A., and Bergherr, E. (2021). Joint modelling approaches to survival analysis via likelihood-based boosting techniques. *Computational and Mathematical Methods in Medicine*, 2021(1):4384035.
URL: https://doi.org/10.1155/2021/4384035.

Halestrap, A. P. and Meredith, D. (2004). The SLC16 gene family – from monocarboxylate transporters (MCTs) to aromatic amino acid transporters and beyond. *Pflügers Archiv European Journal of Physiology*, 447(5):619–628.
URL: http://dx.doi.org/10.1007/s00424-003-1067-2.

Hans, N., Klein, N., Faschingbauer, F., Schneider, M., and Mayr, A. (2023). Boosting distributional copula regression. *Biometrics*, 79(3):2298–2310.
URL: https://doi.org/10.1111/biom.13765.

He, K., Li, Y., Zhu, J., Liu, H., Lee, J. E., Amos, C. I., Hyslop, T., Jin, J., Lin, H., Wei, Q., and Li, Y. (2016). Component-wise gradient boosting and false discovery control in survival analysis with high-dimensional covariates. *Bioinformatics*, 32(1):50–57.
URL: https://www.doi.org/10.1093/bioinformatics/btv517.

Heller, G. Z. (2024). Simple or complex statistical models: Non-traditional regression models with intuitive interpretations. *Statistical Modelling*, page 1471082X241274405.
URL: https://doi.org/10.1177/1471082X241274405.

Hofner, B., Mayr, A., and Schmid, M. (2016). gamboostLSS: An R package for model building and variable selection in the GAMLSS framework. *Journal of Statistical Software*, 74(1):1–31.
URL: https://doi.org/10.18637/jss.v074.i01.

Hothorn, T., Buehlmann, P., Kneib, T., Schmid, M., and Hofner, B. (2022). *mboost: Model-Based Boosting*. R package version 2.9-7.
URL: https://CRAN.R-project.org/package=mboost.

Ishwaran, H., Kogalur, U. B., Chen, X., and Minn, A. J. (2011). Random survival forests for high-dimensional data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1):115–132.
URL: https://doi.org/10.1002/sam.10103.

Jenssen, T.-K., Kuo, W. P., Stokke, T., and Hovig, E. (2002). Associations between gene expressions in breast cancer and patient survival. *Human Genetics*, 111(4-5):411–420.
URL: https://doi.org/10.1007/s00439-002-0804-5.

Joe, H. (2005). Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419.
URL: https://doi.org/10.1016/j.jmva.2004.06.003.

Klein, J. P. and Moeschberger, M. L. (2003). *Survival Analysis*. Springer New York.
URL: https://doi.org/10.1007/b97377.

Li, X., Wang, C., Wang, S., Hu, Y., Jin, S., Liu, O., Gou, R., Nie, X., Liu, J., and Lin, B. (2021). YWHAE as an HE4 interacting protein can influence the malignant behaviour of ovarian cancer by regulating the PI3K/AKT and MAPK pathways. *Cancer Cell International*, 21(1).
URL: http://dx.doi.org/10.1186/s12935-021-01989-7.

Liu, X.-R., Pawitan, Y., and Clements, M. (2018). Parametric and penalized generalized survival models. *Statistical Methods in Medical Research*, 27(5):1531–1546.
URL: https://doi.org/10.1177/0962280216664760.

Lo, A., Chernoff, H., Zheng, T., and Lo, S.-H. (2015). Why significant variables aren't automatically good predictors. *Proceedings of the National Academy of Sciences*, 112(45):13892–13897.
URL: https://www.doi.org/10.1073/pnas.1518285112.

Marra, G. and Radice, R. (2020). Copula link-based additive models for right-censored event time data. *Journal of the American Statistical Association*, 115(530):886–895.
URL: https://doi.org/10.1080/01621459.2019.1593178.

Marra, G. and Radice, R. (2023). *GJRM: Generalised Joint Regression Modelling*. R package version 0.2-6.4.
URL https://CRAN.R-project.org/package=GJRM.

Mayr, A., Binder, H., Gefeller, O., and Schmid, M. (2014). The evolution of boosting algorithms: From Machine Learning to Statistical Modelling. *Methods of Information in Medicine*, 53(6):419–427.
URL: https://www.doi.org/10.3414/ME13-01-0122.

Mayr, A., Fenske, N., Hofner, B., Kneib, T., and Schmid, M. (2012). Generalized Additive Models for Location, Scale and Shape for high dimensional data — A flexible approach based on Boosting. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 61(3):403–427.
URL: https://doi.org/10.1111/j.1467-9876.2011.01033.x.

Mayr, A., Hofner, B., and Schmid, M. (2016). Boosting the discriminatory power of sparse survival models via optimization of the concordance index and stability selection. *BMC Bioinformatics*, 17(1):288.
URL: https://doi.org/10.1186/s12859-016-1149-8.

Morris, E., He, K., Li, Y., Li, Y., and Kang, J. (2020). SurvBoost: An R Package for High-Dimensional Variable Selection in the Stratified Proportional Hazards Model via Gradient Boosting. *The R Journal*, 12(1):105–117.
URL: https://www.doi.org/10.32614/rj-2020-018.

Nagler, T., Schepsmeier, U., Stoeber, J., Brechmann, E. C., Graeler, B., and Erhardt, T. (2022). *VineCopula: Statistical Inference of Vine Copulas*. R package version 2.4.4.
URL: https://CRAN.R-project.org/package=VineCopula.

Norman, P. A., Li, W., Jiang, W., and Chen, B. E. (2024). deepAFT: A nonlinear accelerated failure time model with artificial neural network. *Statistics in Medicine*, 43(19):3689–3701.
URL: https://doi.org/10.1002/sim.10152.

Othus, M., Barlogie, B., LeBlanc, M. L., and Crowley, J. J. (2012). Cure models as a useful statistical tool for analyzing survival. *Clinical Cancer Research*, 18(14):3731–3736.
URL: https://doi.org/10.1158/1078-0432.CCR-11-2859.

Ovarian Cancer Research Alliance (OCRA) (2021).
URL: https://ocrahope.org/news/high-grade-serous-carcinoma/#:~:text=High%2Dgrade%20serous%20carcinoma%20is,unless%20another%20type%20is%20specified.

Parsa, M., Taghavi-Shahri, S. M., and Van Keilegom, I. (2024). On variable selection in a semiparametric AFT mixture cure model. *Lifetime Data Analysis*, 30(2):472–500.
URL: https://doi.org/10.1007/s10985-024-09619-w.

Peng, Y. and Yu, B. (2021). *Cure Models: Methods, Applications, and Implementation*. Chapman and Hall/CRC.
URL: http://dx.doi.org/10.1201/9780429032301.

Petti, D., Eletti, A., Marra, G., and Radice, R. (2022). Copula link-based additive models for bivariate time-to-event outcomes with general censoring scheme. *Computational Statistics & Data Analysis*, 175:107550.
URL: https://doi.org/10.1016/j.csda.2022.107550.

Popple, A., Durrant, L. G., Spendlove, I., Rolland, P., Scott, I. V., Deen, S., and Ramage, J. M. (2012). The chemokine, CXCL12, is an independent predictor of poor survival in ovarian cancer. *British Journal of Cancer*, 106(7):1306–1313.
URL: http://dx.doi.org/10.1038/bjc.2012.49.

Ren, X., Wang, X., Peng, B., Liang, Q., Cai, Y., Gao, K., Hu, Y., Xu, Z., and Yan, Y. (2021). Significance of TEAD family in diagnosis, prognosis and immune response for Ovarian Serous Carcinoma. *International Journal of General Medicine*, 14:7133–7143.
URL: http://dx.doi.org/10.2147/IJGM.S336602.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized Additive Models for Location, Scale and Shape. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3):507–554.
URL: https://doi.org/10.1111/j.1467-9876.2005.00510.x.

Rizopoulos, D. (2012). *Joint Models for Longitudinal and Time-to-Event Data*. Chapman and Hall/CRC.
URL: https://doi.org/10.1201/b12208.

Salerno, S. and Li, Y. (2023). High-dimensional survival analysis: Methods and applications. *Annual Review of Statistics and Its Application*, 10:25–49.
URL: https://www.doi.org/10.1146/annurev-statistics-032921-022127.

Siegel, R. L., Miller, K. D., and Jemal, A. (2020). Cancer statistics, 2020. *CA: A Cancer Journal for Clinicians*, 70(1):7–30.
URL: https://doi.org/10.3322/caac.21590.

Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S., and Mayr, A. (2022). Deselection of base-learners for Statistical Boosting with an application to distributional regression. *Statistical Methods in Medical Research*, 31(2):207–224.
URL: https://doi.org/10.1177/09622802211051088.

Sun, T. and Ding, Y. (2019). Copula-based semiparametric regression method for bivariate data under general interval censoring. *Biostatistics*, 22(2):315–330.
URL: https://doi.org/10.1093/biostatistics/kxz032.

Tang, X., Qi, C., Zhou, H., and Liu, Y. (2022). Critical roles of PTPN family members regulated by non-coding RNAs in tumorigenesis and immunotherapy. *Frontiers in Oncology*, 12.
URL: http://dx.doi.org/10.3389/fonc.2022.972906.

TCGA Research Network (2024). The Cancer Genome Atlas.
URL: https://www.cancer.gov/tcga.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A., and Hofner, B. (2018). Gradient Boosting for distributional regression: Faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28(3):673–687.
URL: https://doi.org/10.1007/s11222-017-9754-6.

Tibshirani, R. (1997). The LASSO method for variable selection in the Cox model. *Statistics in Medicine*, 16(4):385–395.
URL:https://doi.org/10.1002/(SICI)1097-0258(19970228)16:4<385::AID-SIM380>3.0.CO;2-3.

Tukey, J. W. (1993). Tightening the clinical trial. *Controlled Clinical Trials*, 14(4):266–285.
URL: https://doi.org/10.1016/0197-2456(93)90225-3.

Uffelmann, E., Huang, Q. Q., Munung, N. S., de Vries, J., Okada, Y., Martin, A. R., Martin, H. C., Lappalainen, T., and Posthuma, D. (2021). Genome-wide association studies. *Nature Reviews Methods Primers*, 1(1).
URL: https://doi.org/10.1038/s43586-021-00056-9.

Wang, H. and Li, G. (2017). A selective review on Random Survival Forests for high dimensional data. *Quantitative Bio-Science*, 36(2):85.
URL: https://www.doi.org/10.22283/qbs.2017.36.2.85.

Wang, W. (2003). Estimating the association parameter for copula models under dependent censoring. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 65(1):257–273.
URL: https://doi.org/10.1111/1467-9868.00385.

Wang, Z., Xing, L., Huang, Y., and Han, P. (2022). FAT2 mutation is associated with better prognosis and responsiveness to immunotherapy in uterine corpus endometrial carcinoma. *Cancer Medicine*, 12(3):3797–3811.
URL: http://dx.doi.org/10.1002/cam4.5119.

Wei, Y., Wojtyś, M., Sorrell, L., and Rowe, P. (2023). Bivariate copula regression models for semi-competing risks. *Statistical Methods in Medical Research*, 32(10):1902–1918.
URL: https://www.doi.org/10.1177/09622802231188516.

Wood, S. N. (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC.
URL: http://dx.doi.org/10.1201/9781315370279.

Wu, M., Lou, J., Zhang, S., Chen, X., Huang, L., Sun, R., Huang, P., Pan, S., and Wang, F. (2016). Gene expression profiling of $CD8^+$ T cells induced by ovarian cancer cells suggests a possible mechanism for $CD8^+$ Treg cell production. *Cell Proliferation*, 49(6):669–677.
URL: http://dx.doi.org/10.1111/cpr.12294.

Yu, S., Wu, Y., Li, C., Qu, Z., Lou, G., Guo, X., Ji, J., Li, N., Guo, M., Zhang, M., Lei, L., and Tai, S. (2020). Comprehensive analysis of the SLC16A gene family in pancreatic cancer via integrated bioinformatics. *Scientific Reports*, 10(1).
URL: http://dx.doi.org/10.1038/s41598-020-64356-y.

Zhang, B., Hepp, T., Greven, S., and Bergherr, E. (2022). Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Computational Statistics*, 37(5):2295–2332.
URL: http://dx.doi.org/10.1007/s00180-022-01199-3.

Supplementary Material

for

"Boosting distributional copula regression for bivariate right-censored time-to-event data"

Contents

Part A: Details on implemented marginal distributions and copula functions.

Part B: Details on the boosting algorithm, software implementation and negative gradients of the loss functions.

Part C: Additional results for the simulation study.

Part A

Distribution	$\boldsymbol{\vartheta}$	Survival function
Weibull	ϑ_1, ϑ_2	$\exp\left(-\left(\frac{t}{\vartheta_1}\right)^{\vartheta_2}\right)$
Log-normal	ϑ_1, ϑ_2	$1 - \Phi\left(\frac{\log(t)-\vartheta_1}{\vartheta_2}\right)$
Log-logistic	ϑ_1, ϑ_2	$1 - \frac{1}{\left(1+\left(\frac{t}{\vartheta_1}\right)^{-\vartheta_2}\right)}$

Table A1: Implemented parametric distributions for right-censored time-to-event responses in `gamboostLSS`. All distribution parameters use the exponential response function, i.e. $\vartheta = \exp(\eta) \geq 0$ except for ϑ_1 in the Log-normal distribution, which uses the identity link function, i.e. $\vartheta = \eta \in \mathbb{R}$.

Table A2: Details of implemented copulas for right-censored time-to-event responses. The functions $\Phi_1^{-1}(\cdot)$ and $\Phi_2(\cdot)$ denote the quantile function and CDF of the univariate and bivariate standard normal distributions, respectively. Rotated copulas by 90, 180 and 270 degrees are respectively defined as: $C_{90} = S_2 - C(1-S_1, S_2; \vartheta^{(c)})$, $C_{180} = S_1 + S_2 - 1 + C(1-S_1, 1-S_2; \vartheta^{(c)})$ and $C_{270} = S_1 - C(S_1, 1-S_2; \vartheta^{(c)})$. The term $D_1(\vartheta^{(c)}) = \int_0^{\vartheta^{(c)}} \frac{t}{\exp(t)-1} dt$ is the Debye function and Φ_2 denotes the CDF of the bivariate Gaussian distribution with correlation coefficient $\vartheta^{(c)}$.

Copula	$C(S_1, S_2; \vartheta^{(c)})$	Range of $\vartheta^{(c)}$	Link	Kendall's τ
Gauss	$\Phi_2(\Phi_1^{-1}(S_1), \Phi_1^{-1}(S_2); \vartheta^{(c)})$	$\vartheta^{(c)} \in [-1, 1]$	$\tanh^{-1}(\vartheta^{(c)})$	$\frac{2}{\pi}\arcsin(\vartheta^{(c)})$
Clayton	$(S_1^{-\vartheta^{(c)}} + S_2^{-\vartheta^{(c)}} - 1)^{-1/\vartheta^{(c)}}$	$\vartheta^{(c)} \in (0, \infty)$	$\log(\vartheta^{(c)})$	$\frac{\vartheta^{(c)}}{\vartheta^{(c)}+2}$
Gumbel	$\exp\left[-\left\{(-\log(S_1))^{\vartheta^{(c)}} + (-\log(S_2))^{\vartheta^{(c)}}\right\}^{\frac{1}{\vartheta^{(c)}}}\right]$	$\vartheta^{(c)} \in [1, \infty)$	$\log(\vartheta^{(c)} - 1)$	$1 - \frac{1}{\vartheta^{(c)}}$
Joe	$1 - ((1-S_1)^{\vartheta^{(c)}} + (1-S_2)^{\vartheta^{(c)}} - (1-S_1)^{\vartheta^{(c)}}(1-S_2)^{\vartheta^{(c)}})^{(1/\vartheta^{(c)})}$	$\vartheta^{(c)} \in [1, \infty)$	$\log(\vartheta^{(c)} - 1)$	$1 + \frac{4}{\vartheta^{(c)2}} \int_0^1 x\log(x)(1-x)^{2(1-\vartheta^{(c)})/\vartheta^{(c)}} dx$
Frank	$-\vartheta^{(c)^{-1}} \log\Big(1 + (\exp(-\vartheta^{(c)}S_1) - 1)\cdot (\exp(-\vartheta^{(c)}S_2) - 1)/(\exp(-\vartheta^{(c)}) - 1)\Big)$	$\vartheta^{(c)} \in \mathbb{R} \setminus \{0\}$	$\vartheta^{(c)}$	$1 - \frac{4}{\vartheta^{(c)}}[1 - D_1(\vartheta^{(c)})]$

Part B

Algorithm B1 Two-stage, non-cyclic boosting for distributional copula regression of time-to-event responses with faster tuning of fitting iterations $\mathtt{m_{stop}}$ by means of out-of-bag (*oobag*) risk.

Require:

Define the base-learners $b_r^{(\bullet)}(x_r)$ for $r = 1, \ldots, P_k^{(\bullet)}$, $\bullet = 1, 2, c$.

Set the step-length $\mathtt{s_{step}} \ll 1$ as well as the (non-optimal) number of fitting iterations $\mathtt{m}_{\mathtt{stop}}^{(\bullet)}, \bullet = \{1, 2, c\}$.

Set weights indicating the training and $\mathtt{m_{stop}}$-tuning partitions of the sample $n_{\mathtt{train}}$, $n_{\mathtt{mstop}}$.

Set stabilisation to be applied to the negative gradient vector (L_2, median absolute deviation or none).

for $\bullet = \{1, 2\}$ **do**

(1) Initialise all predictors $\hat{\eta}_k^{(\bullet)}$ corresponding to $\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}^{(\bullet)}$ with offset values $\hat{\eta}_{k,[0]}^{(\bullet)}$.

for $m = 1, \ldots, \mathtt{m}_{\mathtt{stop}}^{(\bullet)}$ **do**

for $k = 1, \ldots, K_\bullet$ in $\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}^{(\bullet)}$ **do**

(a) Evaluate the parameter-specific negative gradient vector $-\boldsymbol{g}_{k,[m]}^{(\bullet)}$

$$-\boldsymbol{g}_{k,[m]}^{(\bullet)} = \left(-\boldsymbol{g}_{k,[m]}^{(\bullet)}(\boldsymbol{x}_i)\right)_{i=1,\ldots,n_{\mathtt{train}}} = -\left(\frac{\partial \omega\left(\boldsymbol{y}_i, \hat{\boldsymbol{\eta}}_i^{(\bullet)}\right)}{\partial \eta_k^{(\bullet)}}\bigg|_{\hat{\boldsymbol{\eta}}^{(\bullet)} = \hat{\boldsymbol{\eta}}_{[m-1]}^{(\bullet)}(\boldsymbol{x}_i)}\right)_{i=1,\ldots,n_{\mathtt{train}}}.$$

(b) Fit $-\boldsymbol{g}_{k,[m]}^{(\bullet)}$ to each parameter-specific base-learner $b_{k,j}^{(\bullet)}(x_j)$.

(c) Select the best-fitting base-learner $\hat{b}_{k,j^\star}^{(\bullet)}$ via residual sum of squares criterion.

$$j^\star = \underset{j \in 1, \ldots P_k^{(\bullet)}}{\arg\min} \sum_{i=1}^{n_{\mathtt{train}}} \left(-g_{k,[m]}^{(\bullet)}(\boldsymbol{x}_i) - \hat{b}_{k,j}^{(\bullet)}(x_i)\right)^2.$$

(d) Compute loss reduction of a weak update using $\hat{b}_{k,j^\star}^{(\bullet)}$.

$$\Delta\omega_{\vartheta_k^{(\bullet)}} = \sum_{i=1}^{n_{\mathtt{train}}} \omega\left(\boldsymbol{y}_i; \hat{\eta}_k^{(\bullet)} + \mathtt{s_{step}} \hat{b}_{k,j^\star}^{(\bullet)}(x_{ij^\star})\right).$$

end for

(2) Update the parameter with highest loss reduction $\vartheta_k^{(\bullet)^\star} = \arg\min_{\vartheta_k^{(\bullet)} \in \boldsymbol{\vartheta}} \left(\Delta\omega_{\vartheta_k^{(\bullet)}}\right)$:

$$\hat{\eta}_{k,[m]}^{(\bullet)*}(\boldsymbol{x}_i) = \hat{\eta}_{k,[m-1]}^{(\bullet)*}(\boldsymbol{x}_i) + \mathtt{s_{step}} \cdot \hat{b}_{k,j^\star}^{(\bullet)}(x_{ij^\star}).$$

(3) For the remaining parameters $\vartheta_k^{(\bullet)} \neq \vartheta_k^{(\bullet)^\star}$, set $\hat{\eta}_{k,[m]}^{(\bullet)}(\boldsymbol{x}_i) = \hat{\eta}_{k,[m-1]}^{(\bullet)}(\boldsymbol{x}_i)$.

(4) Compute the out-of-bag risk at iteration $[m]$:

$$\mathrm{risk}_{\mathrm{oobag},[m]}^{(\bullet)} = \sum_{i=1}^{n_{\mathtt{mstop}}} \hat{\omega}\left(\boldsymbol{y}_i; \hat{\boldsymbol{\eta}}_i^{(\bullet)}\Big|_{\hat{\eta}^{(\bullet)} = \hat{\eta}_{[m]}^{(\bullet)}(\boldsymbol{x}_i)}\right).$$

end for

(5) Determine $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(\bullet)}$ by means of the out-of-bag-risk:

$$\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(\bullet)} = \underset{m \in 1, \ldots, \mathtt{m}_{\mathtt{stop}}^{(\bullet)}}{\arg\min}\ \mathrm{risk}_{\mathrm{oobag},[m]}.$$

end for

(6) Compute $\hat{S}_\bullet\left(y_{\bullet i}; \hat{\boldsymbol{\vartheta}}_i^{(\bullet)}\right)$, $\hat{f}_\bullet\left(y_{\bullet i}; \hat{\boldsymbol{\vartheta}}_i^{(\bullet)}\right)$ using $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(\bullet)}$, $\bullet = \{1, 2\}$. Plug them into the loss of Equation (3).

(7) Conduct steps (1)-(5) using the loss of Equation (3) with $\bullet = c$ in order to determine $\mathtt{m}_{\mathtt{stop}}^{\mathtt{opt}(c)}$.

Note that during the first for-loop the loss function in steps (1)-(5) in Algorithm B1 is set to the negative log-likelihood of univariate right-censored responses

$$\omega_i = -\ell_i = -\left\{\delta_{\bullet i} \log\left[f_\bullet(y_{\bullet i}; \boldsymbol{\vartheta}_i^{(\bullet)})\right] + (1 - \delta_{\bullet i}) \log\left[S_\bullet(y_{\bullet i}; \boldsymbol{\vartheta}_i^{(\bullet)})\right]\right\}, \quad \bullet = \{1, 2\},$$

whereas for the remainder of the steps it is set to the negative log-likelihood that corresponds to Equation (3).

B1 Fitting bivariate distributional copula regression models for right-censored data using `SurvCopBoost` in R

We briefly illustrate how to use the R routine `SurvCopBoost` which implements Algorithm B1. The function uses syntax similar to that of `mboost`, `gamboostLSS` and other regression routines:

```
## All covariates enter the model of margin 1
Formula_Margin1 <- list(mu = cbind(time1, cens1) ~ .,
                        sigma = cbind(time1, cens1) ~ .)

## All covariates enter the model of margin 2
Formula_Margin2 <- list(mu = cbind(time1, cens1) ~ .,
                        sigma = cbind(time1,cens1) ~ .)

## All covariates enter the model of the copula parameter
Dependence_Formula <- cbind(SURV1, PDF1, delta1,
                            SURV2, PDF2, delta2) ~ .

## Construct list of formulas
formula_list <- list(Formula_Margin1,
                     Formula_Margin2,
                     Dependence_Formula)

## Fit the model, consider 1000 iterations for each sub-model
Fit <- SurvCopBoost(formulas = formula_list,
                    marings = c("WEIBULL", "LOGLOGISTIC"),
                    copula = c("GUMBEL"),
                    response_1 = resp1, response_2 = resp2, data = dat,
                    mstops = c(1000, 1000, 1000),
                    oobag_weights = boost_weights,
                    s_step = 0.1, stabilization = "L2")
```

The argument `formulas` requires a list with three entries that indicate the formulas used for fitting the model of the two margins as well as the dependence parameter $\vartheta^{(c)}$. The marginal distributions are specified in the argument `margins`, which supports the entries `WEIBULL`, `LOGNORMAL`, and `LOGLOGISTIC`. The copula function is determined by the argument `copula`. Rotated copulas are specified by entering the degrees of rotation, e.g. `GUMBEL270` for a Gumbel copula by 270°. The arguments `response_1` and `response_2` are data frames of dimension $n \times 2$, where the first column is the time variable and the second column is the censoring indicator parsed as a binary variable. The explanatory variables are provided in the `data` argument. Note that `data` should not contain the time variables and censoring indicators. A vector of length n consisting only of binary entries must be supplied for `oobag_weights`. This determines the observations used for fitting and for the tuning of `m_stop`. The out-of-bag risk is computed on the observations with weight equal to zero. Lastly, the arguments `mstops`, `s_step` and `stabilization` specify the hyperparameters of the boosting algorithm.

The formula of the dependence parameter declared in `Dependence_Formula` requires the structure with the provided names (`SURV1`, `PDF1`, `delta1`, etc.). These objects denote the survival function, probability density function and the censoring indicator of each margin, respectively. The marginal survival functions and probability density functions are computed internally after boosting each margin as described in Step (6) of Algorithm B1. The output of `SurvCopBoost` is a list which contains the individual sub-models of the margins and the dependence parameter. These objects can then be used with typical convenience functions such as `predict`, `plot`, `coef`, and `summary` from the `gamboostLSS` package.

B2 Negative gradients of the implemented loss functions

Recall that the loss function corresponds to the negative log-likelihood, i.e. $\omega_i = -\ell_i$. The negative gradients of the loss with respect to the additive predictors are then the first partial derivatives of the log-likelihood with respect to the additive predictors, i.e. $-\partial\omega/\partial\eta_{ik}^{(\bullet)} = -\partial(-\ell_i)/\partial\eta_{ik}^{(\bullet)} = \partial\ell_i/\partial\eta_{ik}^{(\bullet)}$, with $\bullet \in \{1, 2, c\}$.

Univariate right-censored time-to-event responses The log-likelihood function for these type of responses is given by:

$$\ell_i = \delta_{\bullet i} \log\{f_{\bullet i}\} + (1 - \delta_{\bullet i}) \log\{S_{\bullet i}\}, \quad \bullet = \{1, 2\}.$$

where the marginal PDFs and survival functions have been abbreviated as $f_{\bullet i} = f_\bullet(y_{\bullet i}; \boldsymbol{\vartheta}_i^{(\bullet)})$ and $S_{\bullet i} = S_\bullet(y_{\bullet i}; \boldsymbol{\vartheta}_i^{(\bullet)})$, $\bullet = \{1, 2\}$, respectively. The negative gradient of the log-likelihood for univariate data is given by:

$$\frac{\partial\ell_i}{\partial\eta_{ik}^{(\bullet)}} = \frac{\delta_{\bullet i}}{f_{\bullet i}} \frac{\partial f_{\bullet i}}{\partial\vartheta_{ik}^{(\bullet)}} \frac{\partial\vartheta_{ik}^{(\bullet)}}{\partial\eta_{ik}^{(\bullet)}} + \frac{(1 - \delta_{\bullet i})}{S_{\bullet i}} \frac{\partial S_{\bullet i}}{\partial\vartheta_{ik}^{(\bullet)}} \frac{\partial\vartheta_{ik}^{(\bullet)}}{\partial\eta_{ik}^{(\bullet)}}, \text{ with } k = 1, \ldots, K_{(\bullet)}, \ \bullet = \{1, 2\},$$

Bivariate right-censored time-to-event responses The log-likelihood function is in this case:

$$\begin{aligned}\ell_i = (1-\delta_{1i})(1-\delta_{2i})\Big\{ &\log(C[S_{1i},\ S_{2i};\ \vartheta^{(c)}])\Big\}+\\ (1-\delta_{1i})\delta_{2i}&\left\{\log\left(\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{2i}}\right)\ +\ \log(f_{2i})\right\}+\\ \delta_{1i}(1-\delta_{2i})&\left\{\log\left(\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_1(y_1;\boldsymbol{\vartheta}^{(1)})}\right)\ +\ \log(f_{1i})\right\}+\\ \delta_{1i}\delta_{2i}&\Big\{\log(c[S_{1i},\ S_{2i};\ \vartheta^{(c)}])+\log(f_{1i})+\log(f_{2i})\Big\},\end{aligned}$$

where once again the marginal PDFs and survival functions have been abbreviated to avoid clutter in the notation and $c[S_{1i},\ S_{2i};\ \vartheta^{(c)}]$ denotes the copula density. The negative gradients are then:

$$\begin{aligned}\frac{\partial \ell_i}{\partial \eta_i^{(c)}} = (1-\delta_{1i})(1-\delta_{2i})&\left\{\frac{1}{C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}}\right\}+\\ (1-\delta_{1i})\delta_{2i}&\left\{\frac{1}{\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{2i}}}\frac{\partial^2 C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{2i}\ \partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}}\right\}+\\ \delta_{1i}(1-\delta_{2i})&\left\{\frac{1}{\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}}}\frac{\partial^2 C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}\ \partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}}\right\}+\\ \delta_{1i}\delta_{2i}&\left\{\frac{1}{c[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}\frac{\partial c[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial \vartheta_i^{(c)}}\frac{\partial \vartheta_i^{(c)}}{\partial \eta_i^{(c)}}\right\},\end{aligned}\tag{SB1}$$

$$\begin{aligned}\frac{\partial \ell_i}{\partial \eta_{ik}^{(1)}} = (1-\delta_{1i})(1-\delta_{2i})&\left\{\frac{1}{C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}}\frac{\partial S_{1i}}{\partial \eta_{ik}^{(1)}}\right\}+\\ (1-\delta_{1i})\delta_{2i}&\left\{\frac{1}{\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{2i}}}\frac{\partial^2 C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{2i}\ \partial S_{1i}}\frac{\partial S_{1i}}{\partial \eta_{ik}^{(1)}}\right\}+\\ \delta_{1i}(1-\delta_{2i})&\left\{\frac{1}{\frac{\partial C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}}}\frac{\partial^2 C[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}^2}\frac{\partial S_{1i}}{\partial \eta_{ik}^{(1)}}\ +\ \frac{1}{f_{1i}}\frac{\partial f_{1i}}{\partial \eta_{ik}^{(1)}}\right\}+\\ \delta_{1i}\delta_{2i}&\left\{\frac{1}{c[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}\frac{\partial c[S_{1i},\ S_{2i};\ \vartheta^{(c)}]}{\partial S_{1i}}\frac{\partial S_{1i}}{\partial \eta_{ik}^{(1)}}\ +\ \frac{1}{f_{1i}}\frac{\partial f_{1i}}{\partial \eta_{ik}^{(1)}}\right\},\end{aligned}\tag{SB2}$$

where $\partial S_{1i}/\partial \eta_{ik}^{(1)} = (\partial S_{1i}/\partial \vartheta_{ik}^{(1)})\ (\partial \vartheta_{ik}^{(1)}/\partial \eta_{ik}^{(1)})$ and $\partial f_{1i}/\partial \eta_{ik}^{(1)} = (\partial f_{1i}/\partial \vartheta_{ik}^{(1)})\ (\partial \vartheta_{ik}^{(1)}/\partial \eta_{ik}^{(1)})$. We remark that BTE estimation requires the negative gradient of Equation (SB1), whereas SCR estimation requires Equation (SB1) and Equation (SB2).

Part C

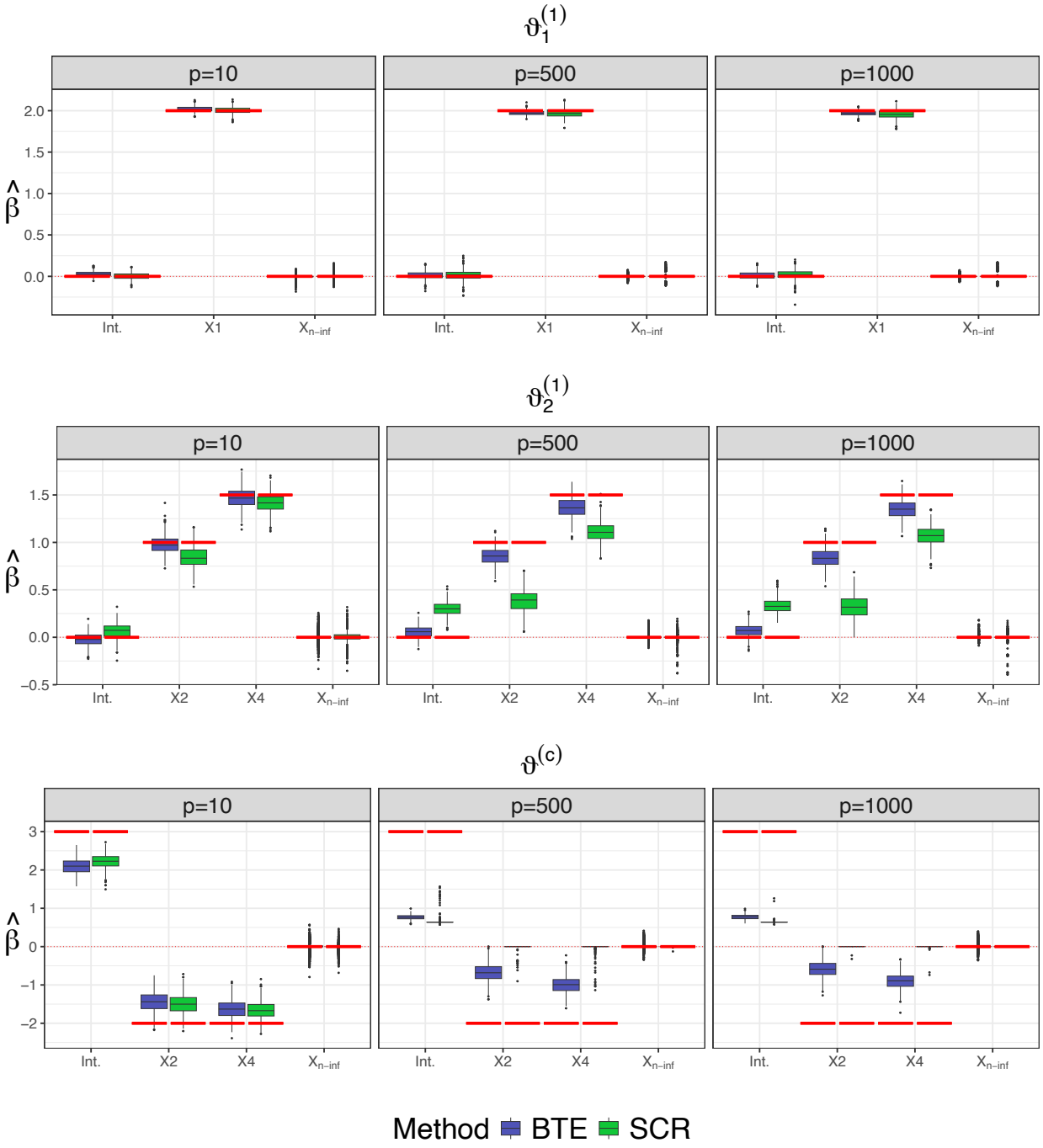

Figure C1: Simulation study 1 (SCR responses). Estimated coefficients of the copula model across distribution parameters, number of potential covariates using BTE and SCR estimation methods based on 500 independent replications. Thick red lines denote true values.

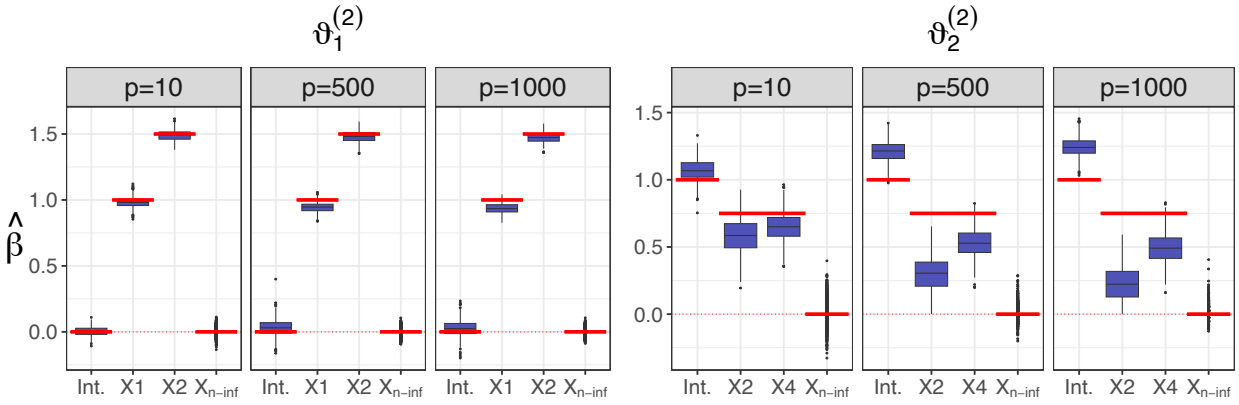

Figure C2: Simulation study 1 (SCR responses). Estimated coefficients of the copula model in the margin corresponding to the terminal event (T_2) across distribution parameters and number of potential covariates using 500 independent replications.

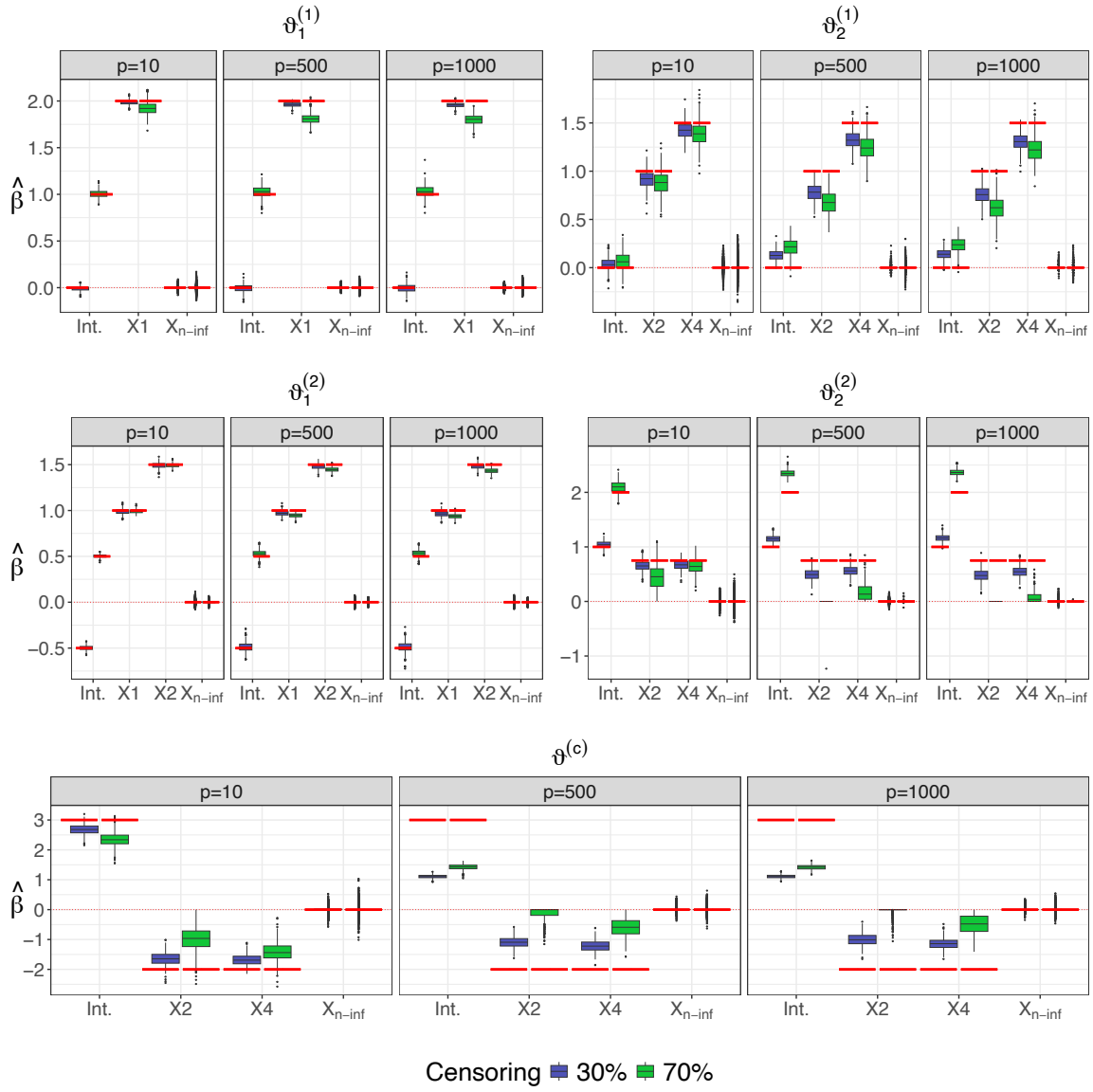

Figure C3: Simulation study 2 linear DGP. Estimated coefficients of the copula model across distribution parameters, number of potential covariates and censoring rates using 500 independent replications. Thick red lines denote true values.

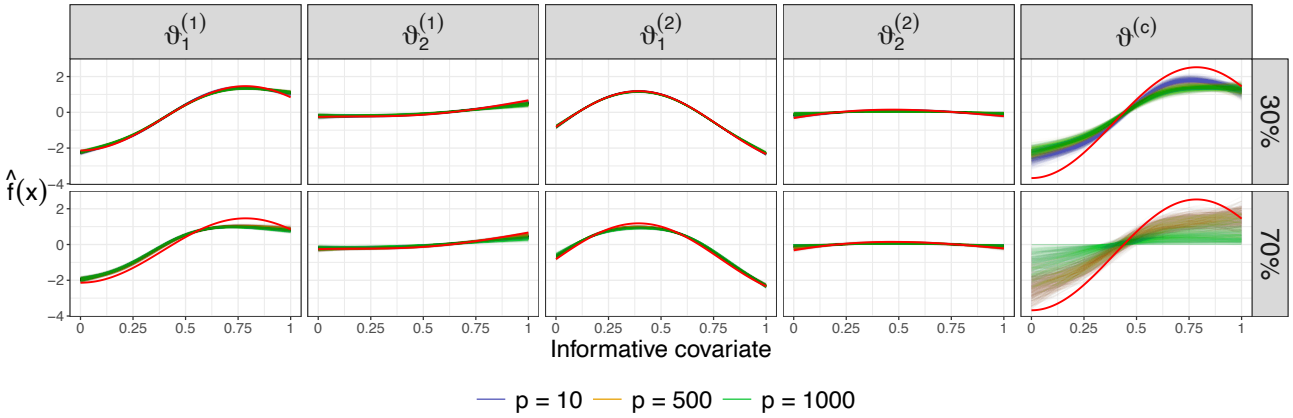

Figure C4: Simulation study 2 non-linear DGP. Estimated non-linear effects of the copula model across distribution parameters, number of potential covariates and censoring rates using 500 independent replications. Thick red lines denote the true non-linear functions.

Table C1: Simulation study 1 (SCR responses). Performance metrics for the simulation studies for the copula models using BTE and SCR estimation, as well as independent univariate Cox models (Cox). Values are mean scores from the 500 independent replicates (each evaluated on the test dataset), whereas parentheses show the respective standard deviations.

	Model	$p_1 = 10$	$p_2 = 500$	$p_3 = 1000$
log-score	BTE	842.821 (38.263)	884.854 (40.024)	894.886 (39.025)
	SCR	829.837 (37.451)	932.535 (35.847)	939.632 (37.148)
	Ind	1257.613 (43.103)	1299.857 (45.343)	1312.84 (43.517)
IBS (T_1)	BTE	0.180 (0.205)	0.182 (0.213)	0.175 (0.195)
	SCR	0.179 (0.203)	0.188 (0.219)	0.181 (0.199)
	Cox	0.458 (0.142)	0.465 (0.147)	0.458 (0.137)
IBS (T_2)	BTE	0.198 (0.231)	0.190 (0.225)	0.182 (0.214)
	SCR	0.198 (0.231)	0.190 (0.225)	0.182 (0.214)
	Cox	0.376 (0.118)	0.371 (0.116)	0.354 (0.109)
ISE (T_1)	BTE	0.002 (0.001)	0.004 (0.001)	0.005 (0.002)
	SCR	0.003 (0.001)	0.015 (0.004)	0.017 (0.005)
	Cox	1.966 (0.127)	1.979 (0.167)	1.979 (0.155)
ISE (T_2)	BTE	0.003 (0.001)	0.009 (0.002)	0.011 (0.003)
	SCR	0.003 (0.001)	0.009 (0.002)	0.011 (0.003)
	Cox	1.736 (0.094)	1.728 (0.129)	1.731 (0.123)
IAE (T_1)	BTE	0.063 (0.018)	0.085 (0.014)	0.092 (0.015)
	SCR	0.070 (0.018)	0.180 (0.023)	0.199 (0.027)
	Cox	2.921 (0.126)	2.938 (0.157)	2.940 (0.148)
IAE (T_2)	BTE	0.073 (0.018)	0.143 (0.020)	0.160 (0.023)
	SCR	0.073 (0.018)	0.143 (0.020)	0.160 (0.023)
	Cox	2.468 (0.089)	2.459 (0.113)	2.464 (0.107)
C-Index (T_1)	BTE	0.824 (0.008)	0.825 (0.008)	0.824 (0.008)
	SCR	0.824 (0.008)	0.824 (0.008)	0.824 (0.008)
	Cox	0.823 (0.008)	0.825 (0.008)	0.824 (0.008)
C-Index (T_2)	BTE	0.862 (0.008)	0.861 (0.008)	0.861 (0.008)
	SCR	0.862 (0.008)	0.861 (0.008)	0.861 (0.008)
	Cox	0.861 (0.008)	0.861 (0.008)	0.860 (0.008)

Gumbel copula with Kendall's τ with range within [0.187; 0.922].

Gradients stabilised using L_2 norm, step-length $\texttt{s}_{\texttt{step}} = 0.1$. $n_{\text{train}} = 1000$, $n_{\text{test}} = 1000$, $n_{\texttt{mstop}} = 1000$.

Table C2: Simulation study 1 (SCR responses). True positive rates (TPR) and false positive rates (FPR) for the copula models using BTE and SCR estimation for each distribution parameter as well as independent univariate Cox models (Cox) for each margin. Values are averages over 500 independent datasets.

	$p_1 = 10$		$p_2 = 500$		$p_3 = 1000$	
	TPR	FPR	TPR	FPR	TPR	FPR
Copula model (BTE)						
$\vartheta_1^{(1)}$	1	0.296	1	0.029	1	0.015
$\vartheta_2^{(1)}$	1	0.173	1	0.001	1	0.000
$\vartheta_1^{(2)}$	1	0.253	1	0.040	1	0.021
$\vartheta_2^{(2)}$	1	0.353	0.990	0.003	0.970	0.001
$\vartheta^{(c)}$	1	0.291	0.998	0.054	0.997	0.027
Copula model (SCR)						
$\vartheta_1^{(1)}$	1	0.145	1	0.004	1	0.002
$\vartheta_2^{(1)}$	1	0.353	1	0.002	0.996	0.000
$\vartheta_1^{(2)}$	1	0.253	1	0.040	1	0.021
$\vartheta_2^{(2)}$	1	0.353	0.990	0.003	0.970	0.001
$\vartheta^{(c)}$	1	0.141	0.055	0.000	0.009	0.000
Cox models (Cox)						
Margin 1	0.791	0.223	0.475	0.034	0.431	0.021
Margin 2	0.913	0.201	0.751	0.039	0.721	0.025

Gumbel copula with Kendall's τ range within [0.187; 0.922].
Gradients stabilised using L_2 norm, step-length $\texttt{s}_{\texttt{step}} = 0.1$.
$n_{\text{train}} = 1000$, $n_{\text{test}} = 1000$, $n_{\texttt{mstop}} = 1000$.

Table C3: Simulation study 2. Performance metrics for the simulation studies for the copula (*Cop*), independent models (*Ind*), and Cox models (*Cox*), $\star$ identifies the non-linear DGP. Values are mean scores from the 500 independent replicates (each evaluated on the test dataset), whereas parentheses show the respective standard deviations.

		(1)			(2)		
	Model	30% censoring			70% censoring		
		$p_1 = 10$	$p_2 = 500$	$p_3 = 1000$	$p_1 = 10$	$p_2 = 500$	$p_3 = 1000$
log-score	*Cop*	1065.961 (39.221)	1105.877 (40.192)	1112.354 (41.310)	774.550 (35.595)	825.408 (35.945)	837.167 (36.137)
	Ind	1462.413 (40.854)	1476.543 (41.899)	1479.465 (41.955)	922.435 (38.587)	951.135 (38.072)	959.964 (37.356)
	Cop $\star$	1103.680 (46.274)	1155.296 (49.054)	1165.732 (44.215)	166.023 (30.444)	189.372 (30.626)	193.941 (30.857)
	Ind $\star$	1383.551 (49.858)	1395.217 (52.250)	1400.523 (46.327)	183.126 (31.038)	199.275 (31.044)	203.357 (31.048)
IBS (T_1)	*Cop*	0.158 (0.198)	0.162 (0.198)	0.139 (0.172)	0.212 (0.181)	0.209 (0.181)	0.197 (0.168)
	Cox	0.402 (0.157)	0.408 (0.155)	0.388 (0.130)	0.488 (0.151)	0.468 (0.147)	0.462 (0.130)
	Cop $\star$	0.249 (0.254)	0.238 (0.242)	0.243 (0.251)	0.276 (0.243)	0.263 (0.235)	0.260 (0.236)
	Cox $\star$	0.315 (0.287)	0.305 (0.284)	0.315 (0.294)	0.258 (0.253)	0.249 (0.246)	0.245 (0.294)
IBS (T_2)	*Cop*	0.154 (0.220)	0.190 (0.257)	0.183 (0.248)	0.109 (0.188)	0.097 (0.176)	0.107 (0.185)
	Cox	0.464 (0.159)	0.490 (0.182)	0.487 (0.177)	0.443 (0.130)	0.435 (0.131)	0.435 (0.130)
	Cop $\star$	0.249 (0.305)	0.252 (0.306)	0.235 (0.292)	0.187 (0.230)	0.191 (0.226)	0.190 (0.229)
	Cox $\star$	0.397 (0.106)	0.393 (0.106)	0.387 (0.101)	0.245 (0.056)	0.239 (0.059)	0.240 (0.058)
ISE (T_1)	*Cop*	0.002 (0.001)	0.006 (0.002)	0.007 (0.002)	0.005 (0.003)	0.017 (0.005)	0.019 (0.005)
	Cox	2.617 (0.268)	2.639 (0.280)	2.633 (0.302)	1.180 (0.078)	1.125 (0.101)	1.128 (0.302)
	Cop $\star$	0.007 (0.003)	0.014 (0.004)	0.015 (0.004)	0.001 (0.000)	0.003 (0.001)	0.003 (0.001)
	Cox $\star$	0.703 (0.047)	0.692 (0.048)	0.694 (0.047)	0.042 (0.004)	0.041 (0.004)	0.041 (0.047)
ISE (T_2)	*Cop*	0.002 (0.001)	0.005 (0.002)	0.006 (0.002)	0.002 (0.001)	0.016 (0.004)	0.019 (0.004)
	Cox	3.157 (0.297)	3.217 (0.343)	3.199 (0.349)	2.263 (0.073)	2.259 (0.078)	2.257 (0.078)
	Cop $\star$	0.003 (0.001)	0.007 (0.002)	0.007 (0.002)	0.001 (0.000)	0.003 (0.001)	0.003 (0.001)
	Cox $\star$	0.710 (0.057)	0.701 (0.058)	0.705 (0.055)	0.077 (0.008)	0.071 (0.006)	0.071 (0.006)
IAE (T_1)	*Cop*	0.057 (0.017)	0.110 (0.019)	0.120 (0.019)	0.109 (0.032)	0.216 (0.036)	0.233 (0.036)
	Cox	3.865 (0.307)	3.903 (0.306)	3.893 (0.330)	2.045 (0.074)	1.994 (0.096)	1.998 (0.330)
	Cop $\star$	0.130 (0.025)	0.186 (0.028)	0.193 (0.028)	0.024 (0.005)	0.036 (0.006)	0.038 (0.007)
	Cox $\star$	1.685 (0.068)	1.680 (0.070)	1.685 (0.070)	0.164 (0.009)	0.165 (0.009)	0.164 (0.070)
IAE (T_2)	*Cop*	0.058 (0.014)	0.106 (0.016)	0.115 (0.016)	0.052 (0.013)	0.145 (0.021)	0.162 (0.019)
	Cox	4.299 (0.301)	4.359 (0.328)	4.335 (0.338)	2.590 (0.072)	2.586 (0.075)	2.588 (0.075)
	Cop $\star$	0.088 (0.016)	0.133 (0.018)	0.140 (0.018)	0.021 (0.003)	0.037 (0.005)	0.040 (0.006)
	Cox $\star$	1.596 (0.074)	1.589 (0.074)	1.595 (0.069)	0.216 (0.010)	0.213 (0.011)	0.212 (0.010)
C-Index (T_1)	*Cop*	0.822 (0.007)	0.823 (0.007)	0.823 (0.007)	0.836 (0.014)	0.837 (0.013)	0.837 (0.012)
	Cox	0.819 (0.007)	0.823 (0.007)	0.822 (0.007)	0.838 (0.013)	0.838 (0.013)	0.838 (0.012)
	Cop $\star$	0.816 (0.007)	0.816 (0.006)	0.815 (0.006)	0.863 (0.014)	0.863 (0.012)	0.863 (0.013)
	Cox $\star$	0.816 (0.007)	0.816 (0.006)	0.815 (0.006)	0.864 (0.014)	0.863 (0.012)	0.863 (0.013)
C-Index (T_2)	*Cop*	0.855 (0.005)	0.855 (0.006)	0.855 (0.006)	0.948 (0.005)	0.948 (0.006)	0.947 (0.006)
	Cox	0.852 (0.005)	0.854 (0.006)	0.854 (0.006)	0.948 (0.005)	0.948 (0.006)	0.948 (0.005)
	Cop $\star$	0.853 (0.006)	0.853 (0.006)	0.853 (0.006)	0.911 (0.014)	0.911 (0.013)	0.911 (0.014)
	Cox $\star$	0.852 (0.007)	0.853 (0.006)	0.853 (0.006)	0.911 (0.014)	0.910 (0.014)	0.910 (0.015)

Clayton copula with Kendall's τ with range within [0.159; 0.907] in linear DGP, and [0.022; 0.917] in non-linear DGP.

Gradients stabilised using L_2 norm, step-length $\texttt{s}_{\texttt{step}} = 0.1$. $n_{\text{train}} = 1000$, $n_{\text{test}} = 1000$, $n_{\texttt{mstop}} = 1000$.

Table C4: Simulation study 2. True positive rates (TPR) and false positive rates (FPR) for the copula (*Cop*) models for each distribution parameter, as well as independent univariate Cox models (*Cox*) for each margin, $\star$ denotes non-linear DGP. Values are averages over 500 independent datasets.

	(1)						(2)					
	30% censoring						70% censoring					
	$p_1 = 10$		$p_2 = 500$		$p_3 = 1000$		$p_1 = 10$		$p_2 = 500$		$p_3 = 1000$	
	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR	TPR	FPR
Linear DGP												
Copula model (*Cop*)												
$\vartheta_1^{(1)}$	1	0.260	1	0.025	1	0.013	1	0.281	1	0.032	1	0.018
$\vartheta_2^{(1)}$	1	0.164	1	0.001	1	0.000	1	0.190	1	0.003	1	0.001
$\vartheta_1^{(2)}$	1	0.251	1	0.035	1	0.021	1	0.325	1	0.025	1	0.011
$\vartheta_2^{(2)}$	1	0.164	1	0.002	1	0.001	1	0.190	0.394	0.000	0.261	0.000
$\vartheta^{(c)}$	1	0.301	1	0.076	1	0.040	0.993	0.221	0.698	0.021	0.604	0.010
Cox models (*Cox*)												
Margin 1	0.855	0.234	0.528	0.035	0.491	0.025	0.877	0.193	0.799	0.032	0.793	0.020
Margin 2	0.951	0.210	0.888	0.041	0.869	0.028	0.945	0.195	0.741	0.041	0.719	0.024
Non-linear DGP												
Copula model (*Cop* $\star$)												
$\vartheta_1^{(1)}$	1	0.190	1	0.011	1	0.006	1	0.214	1	0.012	1	0.006
$\vartheta_2^{(1)}$	1	0.227	1	0.011	1	0.006	1	0.244	1	0.013	1	0.007
$\vartheta_1^{(2)}$	1	0.264	1	0.019	1	0.009	1	0.269	1	0.016	1	0.008
$\vartheta_2^{(2)}$	1	0.227	0.978	0.005	0.960	0.003	1	0.244	0.696	0.004	0.576	0.002
$\vartheta^{(c)}$	1	0.271	1	0.016	1	0.008	1	0.132	0.994	0.004	0.998	0.001
Independent univariate Cox models (*Cox* $\star$)												
Margin 1	0.759	0.253	0.533	0.042	0.532	0.029	0.893	0.257	0.688	0.039	0.662	0.026
Margin 2	0.897	0.246	0.687	0.052	0.641	0.034	0.834	0.280	0.553	0.051	0.543	0.033

Clayton copula with Kendall's τ with range within [0.159; 0.907] in linear DGP, and [0.022; 0.917] in non-linear DGP.

Gradients stabilised using L_2 norm, step-length $\texttt{s}_{\texttt{step}} = 0.1$. $n_{\text{train}} = 1000$, $n_{\text{test}} = 1000$, $n_{\texttt{mstop}} = 1000$.

Appendix E: Software

This Appendix is dedicated to the software used and developed during the time of this dissertation. First, a brief illustration on how to fit 2SGAMLSS is provided, see Appendix A for more details on the methodology. The aforementioned distributional regression technique is suitable for scenarios where endogeneity due to unmeasured confounders is present. Afterwards, we describe the software modified and created to fit the distributional regression models for non-commensurate responses with one time-to-event margin presented in Appendix B. Lastly, we demonstrate how to use the software implementations of the gradient boosting algorithms for copula-based distributional regression models introduced in Appendices C and D.

E1 Fitting 2SGAMLSS using the `mgcv` and `GJRM` packages

The following code snippet illustrates how to fit 2SGAMLSS using the R packages `mgcv` (Wood, 2023) in tandem with `GJRM` (Marra and Radice, 2023). For this illustration, consider a dataset `dat` which consists of one continuous response (`y`), one endogenous treatment encoded as a binary variable (`x_en`), one continuous instrument (`x_iv`), and two continuous exogenous variables (`x_ex1`, `x_ex2`).

```
# Load packages
library(mgcv)
library(GJRM)

# Construct formula for first stage:
formula_1 <- x_en ~ s(x_iv) + s(x_ex1) + s(x_ex2)

## Fit 1ST stage (binary)
first_stage_model <- gam(formula_1,
                         family = binomial(link = "logit"),
                         data = dat)

### Compute expected value of endogenous variable
expectation_x_en <- predict(first_stage_model, type = "response")
```

```
### Compute residuals
xi_hat <- x_en - expectation_x_en

# attach residuals to data:
dat$xi_hat <- xi_hat

# Construct formula for second stage:
formula_2 <- list(y ~ s(x_en) + s(xi_hat) + s(x_ex1) + s(x_ex2),
                  ~ s(x_en) + s(xi_hat) + s(x_ex1) + s(x_ex2) )

### Fit 2ND stage on a Gaussian distribution:
second_stage_model <- gamlss(formula_2,
                             margin = "N",
                             data = dat)
```

The code snippet fits a GAM on the binary treatment variable `x_en` assuming a Bernoulli distribution. It models the instrument as well as the exogenous variables using smooth functions (`s(x_iv)`). The conditional expectation of `x_en` is computed using the function `predict()` after fitting the first-stage model. Afterwards, the first-stage residuals $\hat{\xi}$ (`xi_hat`) are computed and attached to the data. The second-stage model is fitted as a GAMLSS on a Gaussian distribution including a smooth function of $\hat{\xi}$ (`s(xi_hat)`) in each additive predictor.

In cases where the treatment variable is expressed as a continuous variable, one may fit a GAMLSS to `x_en`. The code snippet below assumes a continuous positive treatment, where a Weibull distribution with two parameters is fitted to `x_en` in the following fashion:

```
# Continuous treatment, assuming a Weibull distribution on x_en:
formula_1 <- list(x_en ~ s(x_iv) + s(x_ex1) + s(x_ex2),
                       ~ s(x_iv) +  s(x_ex1) + s(x_ex2) )

# Fit first stage
first_stage_model <- gamlss(formula_1,
                            margin = "WEI",
                            data = dat)

# compute estimated distribution parameters
```

```
param1_x_en <- exp(predict(first_stage_model, eq=1, type="link"))
param2_x_en <- exp(predict(first_stage_model, eq=2, type="link"))

# compute expectation of weibull distribution:
expectation_x_en <- param1_x_en * gamma(1 + 1/param2_x_en)

# compute first-stage residuals
xi_hat <- x_en - expectation_x_en
```

In this case, the conditional expectation of the Weibull distribution depends on both parameters. This highlights the benefit of using a distributional approach, since both parameters of the Weibull distribution are modelled jointly. The first-stage residuals are then included in the second-stage model as shown in the previous code snippet.

E2 Fitting distributional regression models for mixed responses via the GJRM package

In this subsection we analyse bivariate non-commensurable or mixed outcomes that consist of a non-time-to-event and a time-to-event variable. Several modifications were made to the R package GJRM (version 0.2-5) in order to accommodate the data structure required to fit models for mixed non-time-to-event and time-to-event responses as introduced in Section 2.3 and Appendix B. The main fitting function of the package is gjrm(). It takes various arguments required to initiate the estimation of the regression coefficients. Some of these arguments are similar to other regression modelling functions, e.g. formula, data, and margins (marginal distributions akin to family). In its modified version, the function takes the following additional arguments:

- PAMMdataset: Data in long format obtained from the pammtools (PW) or discSurv (DT) packages.
- PAMMoffset: Offset required for the PW approach. If missing, the DT approach is used instead.
- ListOfIDs: List of rows from the long format data that correspond to each observational unit or individual in the sample. The length of ListOfIDs must be equal to n.

The source file of the modified version is available from the GitHub repository https://github.com/GuilleBriseno/DTPW_DistCopulaReg. The modifications made to the package include changes to the function `pream.wm()`, a routine that ensures that all of the arguments provided to `gjrm()` are correct and supported by the function. In the unmodified `gjrm()`, starting values of the coefficient vectors for the copula-based regression models are obtained by fitting independent GAMs on transformed marginal responses using the Gaussian log-likelihood. The modified version checks if a model for mixed responses was specified and obtains starting values for the coefficients of the hazard of the time-to-event margin by fitting a GAM on the long dataset using either the Poisson (PW) or Bernoulli (DT) log-likelihoods. The generation of starting values for the non-time-to-event margin is left unchanged.

The package `GJRM` stores all functions related to computing marginal CDFs, PDFs and derivatives thereof in specific routines depending on the type of marginal responses. The function `distrHsdiscr()` contains the discrete marginal distributions implemented in the package. Joint likelihoods and cumulative probabilities as well as their first and second order partial derivatives w.r.t. the regression coefficient vector $\boldsymbol{\beta}$ of both DT and PW approaches were added to `distrHsdiscr()`. After computing all necessary quantities from the DT or PW approaches, the following functions may make use of them to compute the log-likelihood, score and hessian of a bivariate copula model: `bprobgHsDiscr()`, `bdiscrdiscr()` and `bdiscrcont()`. These functions implement bivariate copula models for binary & discrete, discrete & discrete and discrete & continuous margins, respectively. Note that in the unmodified `GJRM` package (version 0.2-5) the function `pream.wm()` prevents the utilisation of `bdiscrcont()`. We remark that in the three aforementioned functions, the DT or PW functions enter as the discrete margin. Additional changes were made to `bprobgHsDiscr()`, `bdiscrdiscr()` and `bdiscrcont()` to accommodate the structures of the first and second order partial derivatives of the DT and PW functions that are computed in `distrHsdiscr()`.

In addition, a data pre-processing convenience function `prepare_data()` was created in order to facilitate the preparation of the data. A model for mixed outcomes may be fitted using the modified version of the package in tandem with the convenience function `prepare_data()` in the following fashion:

```
# Pre-processing of data using convenience function:
```

```
data_prepared <- prepare_data(type = "DT",
                              data = short_data,
                              the_intervals = cut_points,
                              DT_timeCol = "Time",
                              DT_eventCol = "Censoring")

# Time-to-event margin formula for DT approach
DT_formula <- formula(deltaj_auxvar ~ s(timeInterval, bs="ps") +
                      x1 + x2)

# Non-time-to-event margin formula
NonTtE_formula <- formula(y_ntte ~ x2 + x3 + s(x4, bs="ps"))

# Dependence parameter formula
Dependence_formula <- formula( ~ x1)

Formulas_list <- list(NonTtE_formula,
                      DT_formula,
                      Dependence_formula)

Model_margins <- c("logit", "PO")

copula_model <- gjrm(Formulas_list,
    data = data_prepared$DataShort,
    BivD = "N",      # Gaussian copula
    margins = Model_margins,
    Model = "B",     # Bivariate model
    ListOfIDs = data_prepared$ListOfIndices,
    PAMM_dataset = data_prepared$DataLong)
```

The routine `prepare_data()` returns a list with three entries:

1. `DataShort`,
2. `DataLong`,
3. `ListOfIndices`.

`DataShort` is the original data with minor modifications in the form of supple-

mentary columns that are required to fit the model using `gjrm()`. `DataLong` is the augmented data according to the provided interval borders `cut_points`. Lastly, `ListOfIndices` is a list of length n and each entry within this list is of length $j(i)$. Therefore, `ListOfIndices` indicates the which rows in the augmented data correspond to the i-th observation.

The argument `BivD` determines the assumed parametric copula, whereas `Model` indicates that a bivariate model is being fitted.

```
## PW CASE
maximum_time_cut <- max(short_data$Time)

data_prepared <- prepare_data(type = "PW",
                          data = short_data,
                          the_intervals = cut_points,
                          PW_formula = Surv(Time, Censoring) ~ .,
                          PW_maximum_time_cut = maximum_time_cut)



# Time-to-event margin formula
PW_formula <- formula(deltaj_auxvar ~ s(tend, bs="ps")+
                      x1 + x2)

# Non-time-to-event margin formula
NonTtE_formula <- formula(y_ntte ~ x2 + x3 +
                           s(x4, bs="ps"))



# Dependence parameter formula
Dependence_formula <- formula( ~ x1 )

Formulas_list <- list(NonTtE_formula,
                      PW_formula,
                      Dependence_formula)

Model_margins <- c("logit", "PO")

copula_model <- gjrm(Formulas_list,
```

```
  data = data_prepared$DataShort,
  BivD = "N",
  margins = Model_margins,
  Model = "B",
  ListOfIDs = data_prepared$ListOfIndices,
  PAMM_dataset = data_prepared$DataLong,
  PAMM_offset = data_prepared$DataLong$offset)
```

The main advantage of relying on the `GJRM` infrastructure is that a model can be fitted to data using a few lines of code. In addition, the user has access to various convenience functions or methods from the package such as `summary()`, `predict()`, `plot()` or `coef()`. In particular, the functions `post.check()` and `conv.check()` provide diagnostic checks for the fitted model, such as convergence of the algorithm, largest absolute value in the score vector, positive semi-definiteness of the Hessian as well as randomised quantile residuals. For the model of the time-to-event margin, we provide an additional helper function `compute_relevant_residuals()` to construct Cox-Snell as well as deviance residuals in order to check the goodness-of-fit of said margin:

```
the_residuals <- compute_relevant_residuals(fitted_model,
                    data_short = data_prepared$Data_Short,
                    data_long = data_prepared$Data_Long,
                    type = "DT",
                    indices = data_prepared$ListOfIndices,
                    equation_survmodel = 2)
```

The convenience function `compute_relevant_residuals()` requires a fitted model, the original data in short format (`data_short`), the long format or augmented data (`data_long`) as well as the list of indices generated by `prepare_data()` (`indices`). The user is able to compute Cox-Snell residuals for the *DT* or *PW* approach by changin the argument `type`. The argument `equation_survmodel` indicates which equation of the `GJRM` object is used for the time-to-event margin. In Appendix B this corresponds to setting `equation_survmodel = 2`. However, when modelling continuous or discrete non-time-to-event margins, this argument may be set to `equation_survmodel=1` due to `GJRM` assigning the first equation to the discrete margin.

It is possible to fit distributional regression models to other combinations of non-commensurable responses with one right-censored time-to-event margin using the

modified version of `GJRM`. Currently the implementation supports non-time-to-event margins that are discrete as well as continuous variables. The code snippet below illustrates how to fit a model to these responses using the *DT* approach to model the hazard rate of the time-to-event margin. It is assumed that the data has already been pre-processed using the `prepare_data()` routine.

```
# Non-time-to-event margin formula
NonTtE_formula <- formula(y_ntte ~ x2 + x3 +
                          s(x4, bs="ps"))

# Formula for second parameter of non-time-to-event margin
NonTtE_formula_2 <- formula(~ x2 + x3)

# Time-to-event margin formula
DT_formula <- formula(deltaj_auxvar ~ s(timeInterval, bs="ps")+
                      x1 + x2)

# Dependence parameter formula
Dependence_formula <- formula( ~ x1)

# Mixed discrete & time-to-event.
# Discrete margin goes first.
Formulas_list <- list(NonTtE_formula,
                      NonTtE_formula_2,
                      DT_formula,
                      Dependence_formula)

# Discrete margin uses Negative Binomial distribution
Model_margins <- c("NBI", "PO")

copula_model <- gjrm(Formulas_list,
  data = data_prepared$DataShort,
  BivD = "N",
  margins = Model_margins,
  Model = "B",
  ListOfIDs = data_prepared$ListOfIndices,
  PAMM_dataset = data_prepared$DataLong)
```

```
# Mixed continuous & time-to-event.
# Continuoues margin must be second equation.
Formulas_list <- list(DT_formula,
                      NonTtE_formula,
                      NonTtE_formula_2,
                      Dependence_formula)

# Continuous margin uses Gaussian distribution
Model_margins <- c("N", "PO")

copula_model <- gjrm(Formulas_list,
  data = data_prepared$DataShort,
  BivD = "N",
  margins = Model_margins,
  Model = "B",
  ListOfIDs = data_prepared$ListOfIndices,
  PAMM_dataset = data_prepared$DataLong)
```

As shown in the snippet above, the implemented modifications easily allow for fitting of various types of non-commensurable outcomes with one right-censored time-to-event margin.

E3 Fitting distributional regression models for bivariate time-to-event responses in GJRM using DT or PW functions

In addition to the three types of non-time-to-event margins (binary, discrete and continuous), the modified GJRM package features an additional implementation of copula models for bivariate right-censored time-to-event responses that use the DT or PW functions as introduced in Section 2.3 and Appendix B. We recommend to define one common grid of cut-off points for the intervals used in the data augmentation for both margins. The code snippet below illustrates how to use the modified R package to fit the model:

```
# Generate data in long format for each margin:
```

```
data_prep_1 <- prepare_data(type = "DT",
                            data = short_data,
                            the_intervals = cut_points,
                            DT_timeCol = "Time1",
                            DT_eventCol = "Censoring1")
data_prep_2 <-  prepare_data(type = "DT",
                             data = short_data,
                             the_intervals = cut_points,
                             DT_timeCol = "Time2",
                             DT_eventCol = "Censoring2")

### Declare model equations
Hazard_1 <- formula(deltaj_auxvar1 ~ s(timeInterval_1, bs="ps")+
                    x1 + x2)
Hazard_2 <- formula(deltaj_auxvar2 ~ s(timeInterval_2, bs="ps")+
                    x1 + x2)
Dependence_formula <- ~ x1

# List of formulas
Formulas_list <- list(Hazard_1,
                      Hazard_2,
                      Dependence_formula)

# Declare margins (must be like this for GJRM to run internally)
the_margins <- c("PO", "PO")

# Fit
copula_model <- gjrm(Formulas_list,
        data = data_prep_1$DataLong,
        BivD = "N",
        margins = Model_margins,
        Model = "B",
        DiscRepresentation = "yes",
        ListOfIDs = data_prep_1$ListOfIndices,
        ListOfIDs_margin2 = data_prep_2$ListOfIndices,
        PAMM_dataset = data_prep_2$DataLong,
```

```
PAMMoffset = NULL, PAMMoffset_2 = NULL)
```

If the arguments `PAMMoffset` and `PAMMoffset_2` are added, the routine will fit the model of the margins using the *PW* approach. Currently, our modified version of `GJRM` supports only the clog-log link function when using the *DT* approach. We plan to add other link functions such as logit and probit in the future. The custom argument `DiscRepresentation = "yes"` allows to fit the model for bivariate time-to-event data.

E4 Boosting distributional copula regression for bivariate binary, discrete and mixed responses using the `gamboostLSS` package

We now describe the software implementation of the boosting algorithm introduced in Appendix C for simultaneous estimation of all model coefficients in detail. The source code for the simultaneous estimation boosting algorithm is available from the following GitHub repository https://github.com/GuilleBriseno/BoostDistCopReg_BinDiscMix. The repository includes several response-type-specific routines to construct loss functions based on bivariate copulas. These loss functions are an extension of the routines introduced in Hans et al. (2023) and are meant to be used in tandem with the `R` package `gamboostLSS`. Supported response types are bivariate binary, bivariate discrete or mixed binary & continuous variables. In addition, as elaborated on Subsection 5.3, further response types are already implemented: mixed binary & discrete and mixed discrete & continuous. Marginal distributions can be specified by the user using the arguments `marg1` and `marg2`. Other arguments that are standard for `gamboostLSS` routines such as `stabilization` and `offset` are supported as well. The code below provides some examples on how to fit each type of bivariate response with some of the implemented copula functions and applying different types of stabilisation:

```
### y1 and y2 are some suitable responses
### All covariates are considered for all parameters
biv_resp_form <- cbind(y1, y2) ~ .

### Bivariate binary responses using Gaussian copula:
```

```
### Probit and logit link functions
Copula_obj <- Gauss_Cop_BivBinary(marg1 = "PROBIT",
                                  marg2 = "LOGIT",
                                  stabilization = "L2")

### Bivariate discrete responses using Joe copula:
### Zero-altered logarithmic and zero-inflated negative binomial
Copula_obj <- Joe_Cop_BivDiscrete(marg1 = "ZALG",
                                  marg2 = "ZINBI",
                                  stabilization = "MAD")

### Mixed binary & continuous responses using rotated
### Clayton copula by 180 degrees:
### Clog-log link and Gaussian margins
Copula_obj <- Clayton180_Cop_BinCont(marg1 = "CLOGLOG",
                                     marg2 = "NORM",
                                     stabilization = "none")

### Mixed binary & discrete responses using Gumbel copula:
### Probit link and zero-inflated negative binomial margins
Copula_obj <- Gumbel_Cop_MixedBinDisc(marg1 = "PROBIT",
                                      marg2 = "ZINBI",
                                      stabilization = "L2")

### Mixed discrete & continuous responses using Frank copula:
### Geometric and Gaussian margins
Copula_obj <- Frank_Cop_MixedDiscCont(marg1 = "GEOM",
                                      marg2 = "NORM",
                                      stabilization = "L2")

### Fit the bivariate responses and boost for 1000 iterations
fit <- gamboostLSS(biv_resp_form,
                   data = data,
                   families = Copula_obj,
                   method = "noncyclic",
                   control = boost_control(mstop = 1000,
```

```
                                        nu = 0.1))
```

Using the infrastructure of `gamboostLSS` offers the benefit of the convenience functions `summary()`, `coef()`, `plot()` or `predict()`. Other functions from the package remain usable.

E5 Boosting distributional copula regression for bivariate right-censored time-to-event data using the gamboostLSS package

We now describe the developed routines for fitting right-censored bivariate time-to-event responses using bivariate copulas in the package `gamboostLSS`. These functions reflect the algorithm introduced in Appendix D which estimates the unknown model coefficients in a two-step fashion. The source code for the two-step estimation boosting algorithm for bivariate time-to-event responses is available from the following GitHub repository https://github.com/GuilleBriseno/BoostDistCopReg_Surv. Using the provided functions, the user is able to construct three loss functions: One for each marginal response and one for the copula. These routines support all arguments of distribution families for the `gamboostLSS` package. The wrapper function `SurvCopBoost` is the main interface to fit distributional copula regression models with automatic data-driven variable selection. This function combines the aforementioned two-step algorithm into one function call in order to streamline model fitting in practice. The `SurvCopBoost` routine uses syntax similar to that of `mboost`, `gamboostLSS` and other regression routines. The code below provides an example on how to fit a model using the developed software:

```
## All covariates enter the model of margin 1
Formula_Margin1 <- list(mu = cbind(time1, cens1) ~ .,
                        sigma = cbind(time1, cens1) ~ .)

## All covariates enter the model of margin 2
Formula_Margin2 <- list(mu = cbind(time2, cens2) ~ .,
                        sigma = cbind(time2, cens2) ~ .)

## All covariates enter the model of the copula parameter
```

```
Dependence_Formula <- cbind(SURV1, PDF1, cens1,
                            SURV2, PDF2, cens2) ~ .

## Construct list of formulas
formula_list <- list(Formula_Margin1,
                     Formula_Margin2,
                     Dependence_Formula)

## Fit the model, consider 1000 iterations for each sub-model
Fit <- SurvCopBoost(formulas = formula_list,
                    marings = c("WEIBULL", "LOGLOGISTIC"),
                    copula = c("GUMBEL"),
                    response_1 = resp1, response_2 = resp2,
                    data = dat,
                    mstops = c(1000, 1000, 1000),
                    oobag_weights = boost_weights,
                    s_step = 0.1, stabilization = "L2")
```

The argument `formulas` requires a list with three entries that indicate the formulas used for fitting the model of the two margins as well as the copula dependence parameter $\vartheta^{(c)}$. The marginal distributions are specified in the argument `margins`, which supports the entries `WEIBULL`, `LOGNORMAL`, and `LOGLOGISTIC`. All of the supported marginal distributions consist of two parameters: `mu` and `sigma`, using the naming convention of `gamboostLSS`. The copula function is determined by the argument `copula`. Rotated copulas are specified by entering the degrees of rotation, e.g. `JOE270` for a Joe copula rotated by 270°. The arguments `response_1` and `response_2` are data frames of dimension $n \times 2$, where the first column is the time variable and the second column is the right-censoring indicator parsed as a binary variable. The explanatory variables are provided in the `data` argument. Note that `data` should not contain the time variables (responses) and censoring indicators. A vector of length n consisting only of binary entries must be supplied for `oobag_weights`. This determines the observations used for fitting and for the tuning of `m_stop`. The out-of-bag risk is computed on the observations with weight equal to zero. Lastly, the arguments `mstops`, `s_step` and `stabilization` specify the hyperparameters of the boosting algorithm.

The formula of the dependence parameter declared in `Dependence_Formula` re-

quires the structure with the provided names, i.e. `SURV1`, `PDF1`, `cens1`, etc. These names denote the survival function, probability density function and the censoring indicator of each margin, respectively. The marginal survival functions and probability density functions are computed internally after boosting each margin as described in Step (6) of Algorithm 1 in Appendix D. The output of `SurvCopBoost` is a list which contains the individual sub-models of the margins and the dependence parameter. These objects can then be used with typical convenience functions such as `predict`, `plot`, `coef`, and `summary` from the `gamboostLSS` package. Below we illustrate the steps that are conducted within `SurvCopBoost`:

```
### Formula of margin 1, 2
form1 <- cbind(time1, cens1) ~ .
form2 <- cbind(time2, cens2) ~ .

### Margin 1 is Weibull and margin 2 is log-normal
distribution_1 <- RightCens_Weibull(stabilization = "L2")
distribution_2 <- RightCens_LogNormal(stabilization = "L2")

### boost for 1000 iterations
model_1 <- gamboostLSS(formula = form_1,
                       data = data,
                       families = distribution_1,
                       control = boost_control(mstop = 1000,
                                               nu = 0.1),
                       method = "noncyclic")

### boost for 1000 iterations
model_2 <- gamboostLSS(formula = form_2,
                      data = data,
                      families = distribution_2,
                      control = boost_control(mstop = 1000,
                                              nu = 0.1),
                      method = "noncyclic")

### Compute estimated distribution parameters
vartheta1_1 <- predict(model_1$mu, type = "response")
```

```
vartheta1_2 <- predict(model_2$sigma, type = "response")

vartheta2_1 <- predict(model_2$mu, type = "response")
vartheta2_2 <- predict(model_2$sigma, type = "response")

### Compute the marginal survival functions and densities
SURV1 <- pweibull(time1, shape=vartheta1_1, scale=vartheta1_2,
               lower.tail = FALSE)
PDF1 <- dweibull(time1, shape=vartheta1_1, scale=vartheta1_2)

SURV2 <- plnorm(time2, meanlog=vartheta2_1, sdlog=vartheta2_2,
             lower.tail = FALSE)
PDF2 <- dlnorm(time2, meanlog=vartheta2_1, sdlog=vartheta2_2)

### Formula of the dependence parameter
dep_form <- cbind(SURV1, PDF1, cens1, SURV2, PDF2, cens2) ~ .

### Specify Clayton copula
copula_family <- BivAFT_ClaytonCop(stabilization = "L2")

### Boost for 1000 iterations using Clayton copula
copula_model <- gamboost(dep_form,
                         family = copula_family,
                         control = boost_control(mstop = 1000,
                                                 nu = 0.1))
```

Once again, implementing our estimation approach in the `gamboostLSS` infrastructure offers the benefit of convenience functions like `summary()`, `coef()`, `plot()` or `predict()`. Other functions from the package remain usable.

Lastly, bivariate time-to-event responses with general censoring scheme (right-, left-, and interval censoring) can be fitted using `SurvCopBoost_GENCENS`. The routine requires the same arguments as `SurvCopBoost`:

```
## All covariates enter the model of margin 1
Formula_Margin1 <- list(
                mu = cbind(time1_left, time1_right, cens1) ~ .,
                sigma = cbind(time1_left, time1_right, cens1) ~ .)
```

```
## All covariates enter the model of margin 2
Formula_Margin2 <- list(
                mu = cbind(time2_left, time2_right, cens2) ~ .,
                sigma = cbind(time2_left, time2_right, cens2) ~ .)

## All covariates enter the model of the copula parameter
Dependence_Formula <- cbind(SURV1_LEFT, SURV1_RIGHT, PDF1, cens1,
                        SURV2_LEFT, SURV2_RIGHT, PDF2, cens2) ~ .

## Construct list of formulas
formula_list <- list(Formula_Margin1,
                     Formula_Margin2,
                     Dependence_Formula)

## Fit the model, consider 1000 iterations for each sub-model
Fit <- SurvCopBoost_GENCENS(formulas = formula_list,
                            marings = c("WEIBULL", "LOGLOGISTIC"),
                            copula = c("CLAYTON"),
                            response_1 = resp1,
                            response_2 = resp2,
                            data = dat,
                            mstops = c(1000, 1000, 1000),
                            oobag_weights = boost_weights,
                            s_step = 0.1, stabilization = "L2")
```

In this case, the main difference compared to `SurvCopBoost` is the number of columns that make up the response variable for the fitting routine. When fitting models for time-to-event data with general censoring schemes, it is required to enter the left and right bounds of the time-to-event response $T = [L, R]$, where L and R denote the left and right bounds of the interval. The censoring indicators, i.e. `cens1` and `cens2` must be numeric variables with values 1, 2, 3 or 4, which indicate whether the i-th observations is uncensored (1), right-censored (2), left-censored (3) or interval-censored (4).

The code below illustrates the steps carried out within `SurvCopBoost_GENCENS`:

```
### Create suitable censoring factor variables
```

```
data_correct_format <- re_format_censoring(data, cens1, cens2)

### Formula of margin 1, 2
form1 <- cbind(time1_left, time1_right, cens1) ~ .
form2 <- cbind(time2_left, time2_right, cens2) ~ .

### Margin 1 is Log-logistic and margin 2 is log-normal
distr_1 <- LogLogisticFamily_GenCens(stabilization="L2")
distr_2 <- LogNormalFamily_GenCens(stabilization="L2")

### boost for 1000 iterations
model_1 <- gamboostLSS(formula = form_1,
                       data = data,
                       families = distr_1,
                       control = boost_control(mstop=1000,
                                               nu = 0.1),
                       method = "noncyclic")

### boost for 1000 iterations
model_2 <- gamboostLSS(formula = form_2,
                         data = data,
                         families = distr_2,
                         control = boost_control(mstop=1000,
                                                 nu = 0.1),
                         method = "noncyclic")

### Compute esitmated distribution parameters
vartheta1_1 <-predict(model_1$mu, type = "response")
vartheta1_2 <-predict(model_2$sigma, type = "response")

vartheta2_1 <-predict(model_2$mu, type = "response")
vartheta2_2 <-predict(model_2$sigma, type = "response")

### Compute the marginal survival functions and densities
S1_LEFT <-pfisk(time1_left, scale=vartheta1_1,
                shape1.a=vartheta1_2, lower.tail = FALSE)
```

```
S1_RIGHT <-pfisk(time1_right, scale=vartheta1_1,
                 shape1.a=vartheta1_2, lower.tail = FALSE)
PDF1 <-dfisk(time1_right, scale = vartheta1_1,
             shape1.a = vartheta1_2)

S2_LEFT <-plnorm(time2_left, meanlog=vartheta2_1,
                  sdlog=vartheta2_2, lower.tail = FALSE)
S2_RIGHT <-plnorm(time2_right, meanlog=vartheta2_1,
                   sdlog=vartheta2_2, lower.tail = FALSE)
PDF2 <-dlnorm(time2_right, meanlog=vartheta2_1,
            sdlog=vartheta2_2)

### Formula of the dependence parameter
dep_form <- cbind(SURV1_LEFT, SURV1_RIGHT, PDF1,
                  SURV2_LEFT, SURV2_RIGHT, PDF2,
                  cens1, cens2) ~ .

### Specify Clayton copula
copula_family<-BivAFT_ClaytonCop_GenCens(stabilization="L2")

### Boost for 1000 iterations using Clayton copula
copula_model <- gamboost(dep_form,
                         family = copula_family,
                         control = boost_control(mstop=1000,
                                                 nu = 0.1))
```

Appendix F: Simulation study of non-commensurable outcomes that include a time-to-event margin

We conduct a simulation study in order to investigate the performance of the proposed function $F(\boldsymbol{\delta})$ introduced in Section 2.3. We are primarily interested in answeing the following questions:

Q1: "Can the underlying baseline hazard rate, baseline cumulative hazard and baseline survival functions of the time-to-event margin be recovered using the proposed functions?"

Q2: "Can the underlying effects of covariates on the respective parameters of the marginal distributions as well as the copula dependence parameter be recovered using the proposed distributional copula regression approach?"

The simulation studies are based on structured additive predictors for all distribution parameters of the non-time-to-event variable, the survival function of the time-to-event margin, as well as the dependence parameter of the parametric copula. The following structured additive predictors are used to generate the bivariate mixed outcome:

$$\begin{aligned}
\eta_{i1}^{(1)} &= \beta_{01}^{(1)} \; + \; \beta_{11}^{(1)} x_{1i} \; + \; \beta_{12}^{(1)} x_{2i}, && \text{(non-time-to-event variable)},\\
\eta_{i2}^{(1)} &= \beta_{02}^{(1)} \; + \; \beta_{12}^{(1)} x_{3i} && \text{(non-time-to-event variable)},\\
\eta_{i}^{(2)} &= s_0^{(2)}(t_i) \; + \; \beta_{1}^{(2)} x_{2i} \; + \; \beta_{2}^{(2)} x_{4i}, && \text{(time-to-event variable)},\\
\eta_{i}^{(c)} &= \beta_{0}^{(c)} \; + \; \beta_{1}^{(c)} x_{5i}, && \text{(copula parameter)},
\end{aligned}$$

where $s_0^{(2)}(t_i)$ denotes the smooth baseline hazard as a function of time t. The effect of a covariate (x_{5i}) in the copula parameter θ leads to various degrees of dependence in either positive or negative directions. We consider cases where the the non-time-to-event margin is set as a count as well as continuous random variable. We use both both PW and DT approaches for Q1 and Q2.

The true baseline functions (hazard, cumulative hazard and survival) used in our simulations are shown in Figure F1(a), (b) and (c), respectively. The functions plotted using red lines correspond to the baseline functions used for simulation scenarios with heavy censoring rates. The hazard rate of the event times used in mild censoring scenarios achieves its maximum at time point $t_{mild}^{max\ haz} = 1.809$,

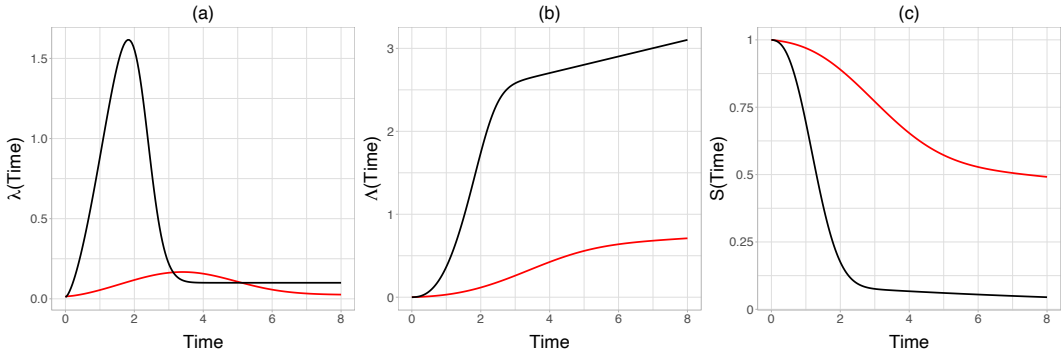

Figure F1: True baseline functions used to generate synthetic data. Black lines correspond to mild censoring, whereas red lines were used for heavy censoring scenarios. Hazard rate (a), cumulative hazard (b), survival function (c).

whereas in the heavy censoring scenarios the mode of the hazard rate is at time point $t_{heavy}^{max\ haz} = 3.417$, see Figure F1(a). Binary margins use only the additive predictor $\eta_{i1}^{(1)}$, whereas two-parameter discrete and continuous distributions use $\eta_{i1}^{(1)}$ and $\eta_{i2}^{(1)}$ for their respective distribution parameters. We set the distributions for discrete and continuous non-time-to-event responses, respectively, to the Negative Binomial Type I (NBI) and Gaussian distributions. We consider the sample sizes $n_1 = 750$, and $n_2 = 1500$. Censoring times are drawn from independent continuous uniform distributions after generating the true event times. The support of the distributions of the censoring times are set such that the censoring rates are approximately 20% (mild censoring case) and 70% (heavy censoring case), respectively.

Bivariate mixed outcomes are generated using the Gaussian, Frank, Clayton and Gumbel copulas. Since the Clayton and Gumbel copulae can only model positive dependence, we employ 90° rotations to generate and fit negative dependence between the observations using these two copulas. We first sample the covariates from independent uniform distributions with support in the interval $[-1, +1]$. Then, we construct the additive predictor for the dependence parameter $(\eta_i^{(c)})$ and sample bivariate uniform margins from a copula with dependence parameter $\vartheta_i^{(c)} = g^{-1}(\eta_i^{(c)})$ using the `R` package `VineCopula`. Afterwards we compute the distribution parameters of the margins and obtain the non-time-to-event response by applying the quantile function of its corresponding distribution to the first sampled uniform margin from the copula. The generation of our time-to-event response is based on Marra and Radice (2020). As a starting point, we construct the survival function using the following predictor:

$$\eta_i^{(2)} = \log[-\log(s_0^{(2)}(t_i))] + \beta_1^{(2)} x_{2i} + \beta_2^{(2)} x_{4i} \Leftrightarrow S(t_i) = \exp[-\exp(\eta_i^{(2)})],$$

Table F1: Overview of scenarios considered in the simulation studies.

Scenario	Non-time-to-event	Censoring rate	n
1	Count (NBI)	Mild ($\approx 20\%$)	750
2	Count (NBI)	Mild ($\approx 20\%$)	1500
3	Continuous (Gauss)	Mild ($\approx 20\%$)	750
4	Continuous (Gauss)	Mild ($\approx 20\%$)	1500
5-8	Add suffix: "`_HEAVY`" for heavy censoring rate ($\approx 70\%$).		
9-16	Add suffix: "`_PW`" for PW approach.		
All fitted using Gaussian, Frank, Clayton, Clayton 90°, Gumbel and Gumbel 90° copulas.			

where $s_0^{(2)}(t_i)$ is the baseline survival function. Then, we numerically invert the second uniform margin sampled from the copula together with the survival function in order to obtain the true event times T_i. The censoring times T_i^{cens} are obtained by sampling from a uniform distribution. The observed times are obtained using $\tilde{T}_i = \min\{T_i, T_i^{cens}\}$ and $\delta_i = \{T_i \leq T_i^{cens}\}$.

In total, $R = 100$ replications are used throughout this section. The number of discrete time intervals is set to $J = 20$ throughout the simulation studies presented in this section, although we remark that setting the number of intervals J to a larger value or to the number of unique event times (e.g. in the PW) case would result in a finer time grid to estimate the hazard rate from. In practice, we recommend to start with a moderate amount of intervals for the time-to-event margin (e.g. 20 intervals) and then increase the number of intervals while assessing the difference of the estimated baseline hazard. We would like to remark that we do not treat the number of intervals as a tuning parameter / hyperparameter in our approach since the regularization induced via the quadratic penalty addresses the compromise between fitting the data too closely and smoothness of the estimated baseline hazard. Table F1 depicts the scenarios considered in our simulation studies.

Results for Q1

The estimated baseline hazard rates using the DT approach are shown in Figures F2 and F3 for scenarios with a binary, discrete and continuous non-time-to-event margin, respectively. Given the chosen configuration for the number of intervals ($J = 20$), the DT approach estimates the underlying baseline hazard

with satisfactory performance. For both mild and heavy censoring cases, it can be seen that the estimated baseline hazard tends to exhibit excessive "wiggliness" or variability towards the end of the study time period, with heavy censoring scenarios showing the most variability in the aforementioned time range. This behaviour is expected, since estimation of any quantity under these circumstances is rather challenging. The additional variability under both censoring regimes is also due to the low number of observations that are observed in that time range. Overall, increasing the sample size leads to a reduced variability in the estimated functions, this can be seen in the blue lines ($n = 1500$) being surrounded by red lines ($n = 750$) in all panels of Figures F2 and F3. Regardless of the type of non-time-to-event margin, the *DT* approach is able to capture the most important aspects of the underlying true baseline hazard rate, these would be for us its shape and (approximate) time point of its mode, see the vertical dotted red lines in the aforementioned figures.

Figures F4 and F5 show the estimated survival functions using the *DT* approach given a binary, discrete and continuous non-time-to-event margin, respectively. Compared to the baseline hazard rate, the estimated baseline survival function using the *DT* approach exhibits a much more robust behaviour. As expected, increasing the sample size leads to a reduced variability in the estimated functions, this reduction can be better appreciated in the panels corresponding to heavy censoring scenarios in Figures F4 and F5. Towards the end of the study time, the fits start to exhibit more variability, especially in cases where heavy censoring is present. As previously mentioned, this behaviour is expected due to a low number of observations in that time range, as well as the high censoring prevalence. Similar to the baseline hazard rate, the estimated baseline survival functions exhibit the same behaviour, regardless of the type of non-time-to-event margin.

As shown in Figures F6, and F7, estimating the hazard of the time-to-event margin using the *PW* approach produces results similar to those of the *DT* approach. Overall, it can be seen that the *PW* approach correctly captures the behaviour of the underlying hazard rate. The estimated hazard rate reaches its mode close to the time point of the true mode. Increasing the sample size reduces the variability in the estimates, as seen from the blue lines being enclosed by the red lines.

Results for Q2.

Figures F10 and F11 show the results of the simulation studies for the *DT* approach using a binary, discrete and continuous non-time-to-event margin, respectively. It can be seen that in general the proposed approach is able to recover the true coefficient values that have an effect on either the respective margins or the dependence parameter with satisfactory performance. The increased variation of the estimated coefficients in the copula dependence parameter increases in cases with heavy censoring, but this behaviour is expected in these challenging scenarios. As expected, whenever a discrete or continuous margin is fitted in the non-time-to-event margin, the estimation of the coefficients improves compared to having a binary non-time-to-event margin. Estimating the coefficients in the dependence parameter seems to be more challenging when using non-elliptical / non-symmetric copulas in cases of negative dependence (i.e. rotations of the Clayton and Gumbel copulae).

The *PW* approach produces results very similar to those obtained using the *DT* approach. Figures F12 and F13 display the bias of the model coefficients across all distribution parameters and fitted copulae. Regardless of the type of considered non-time-to-event variable (binary, count or continuous), the coefficients of all distribution parameters are estimated correctly using elliptical copulas (Gauss and Frank) as well as non-rotated Gumbel and Clayton copulas. Estimation of the effects on the dependence parameter appears to be more challenging for 90 degree rotations of the aforementioned copulas, in particular for the Clayton copula.

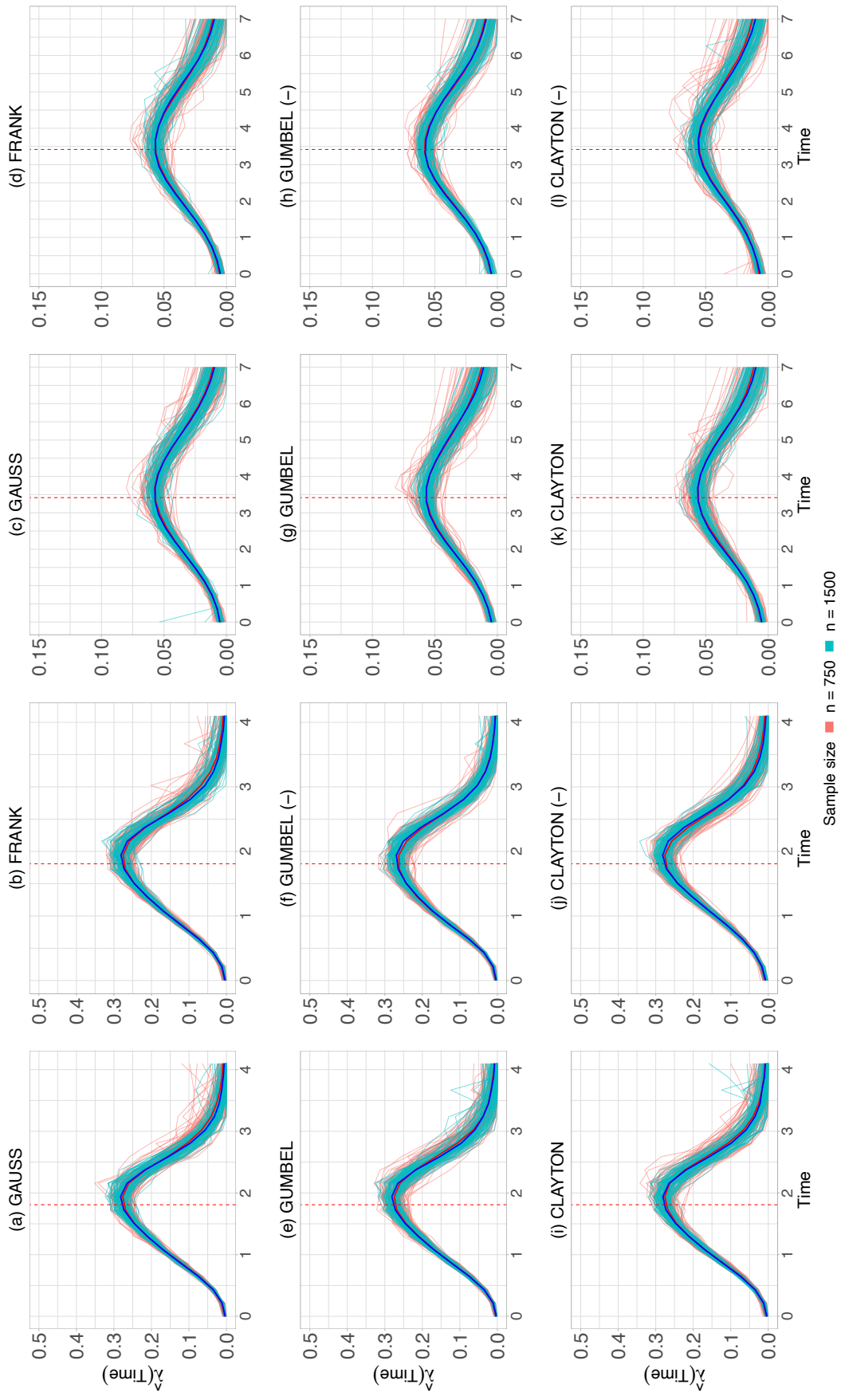

Figure F2: Estimated baseline hazard rate using the *DT* approach across sample sizes and copula functions given a discrete non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $t_{mild}^{max\ haz} = 1.809$ and $t_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

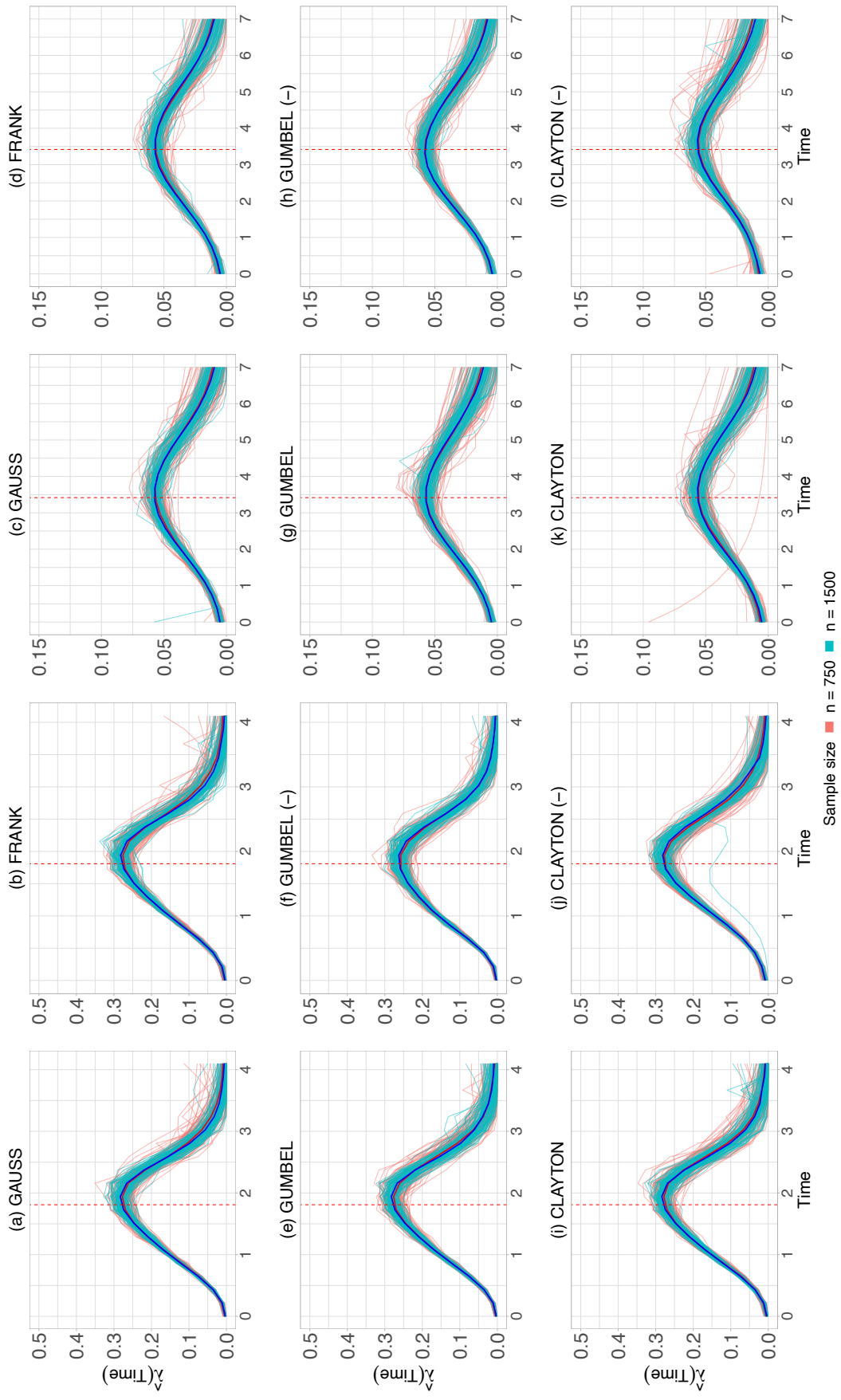

Figure F3: Estimated baseline hazard rate using the *DT* approach across sample sizes and copula functions given a continuous non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $t_{mild}^{max\ haz} = 1.809$ and $t_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

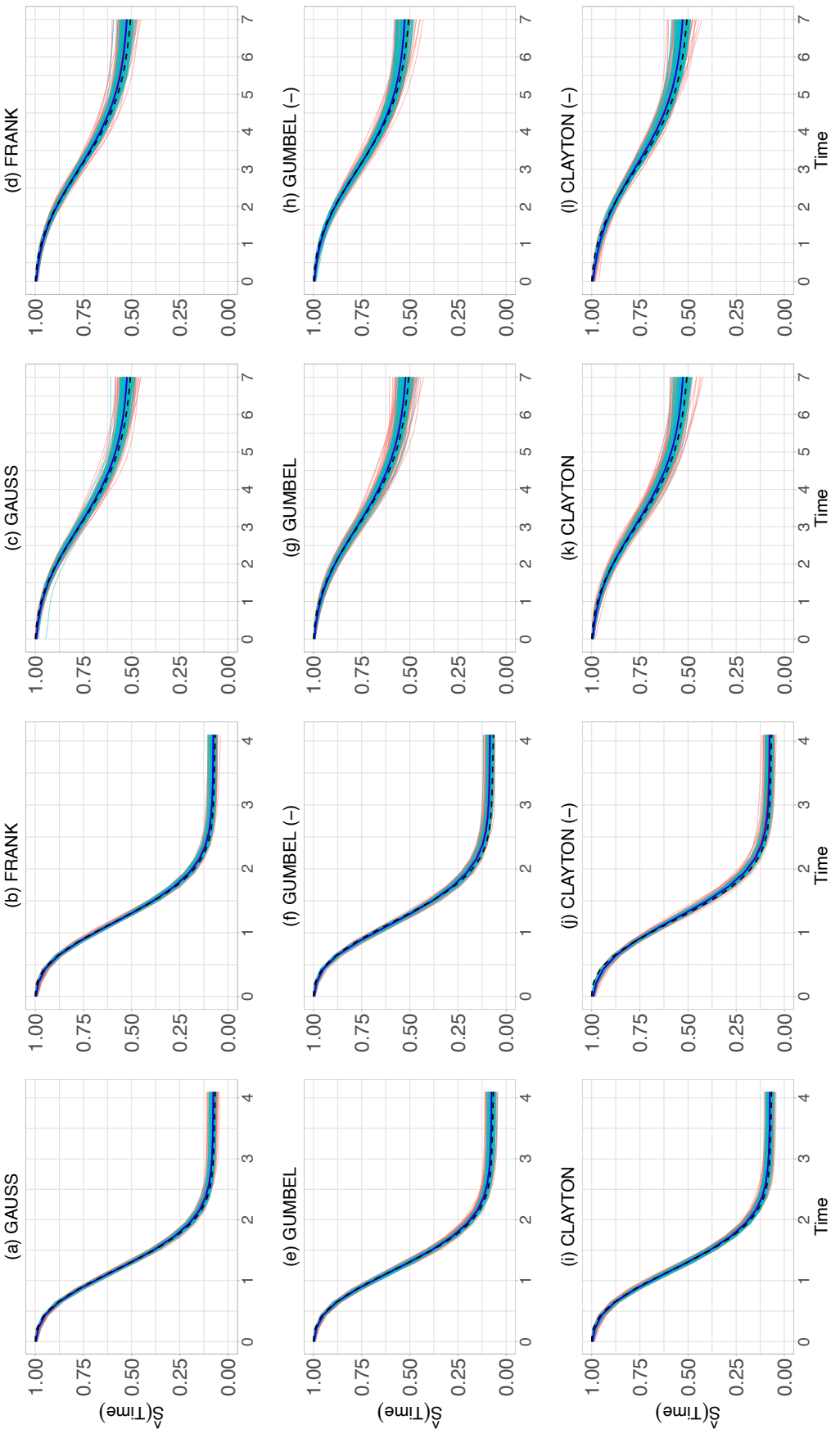

Figure F4: Estimated baseline survival function using the DT approach across sample sizes and copula functions given a discrete non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

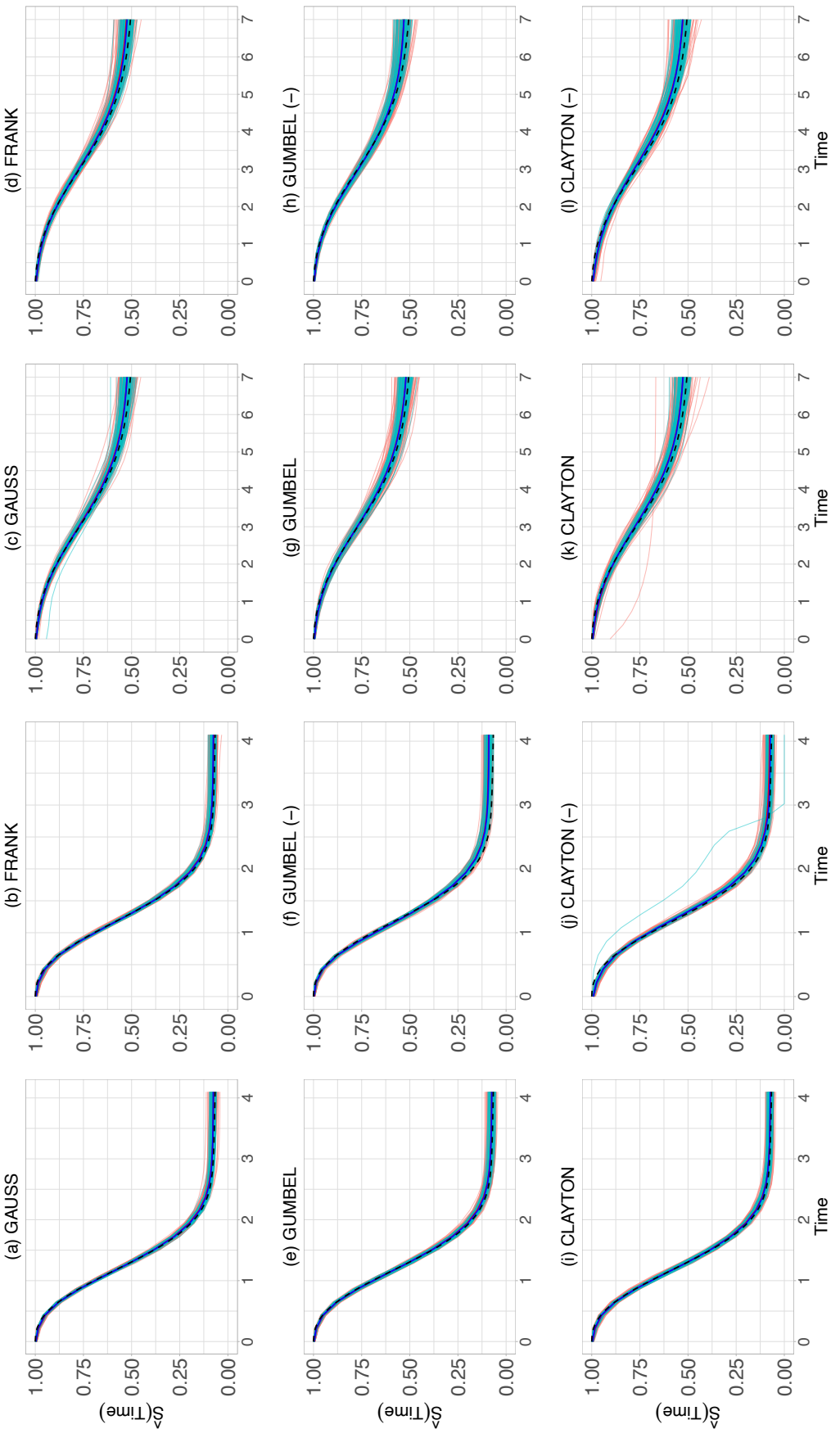

Figure F5: Estimated baseline survival function using the *DT* approach across sample sizes and copula functions given a continuous non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

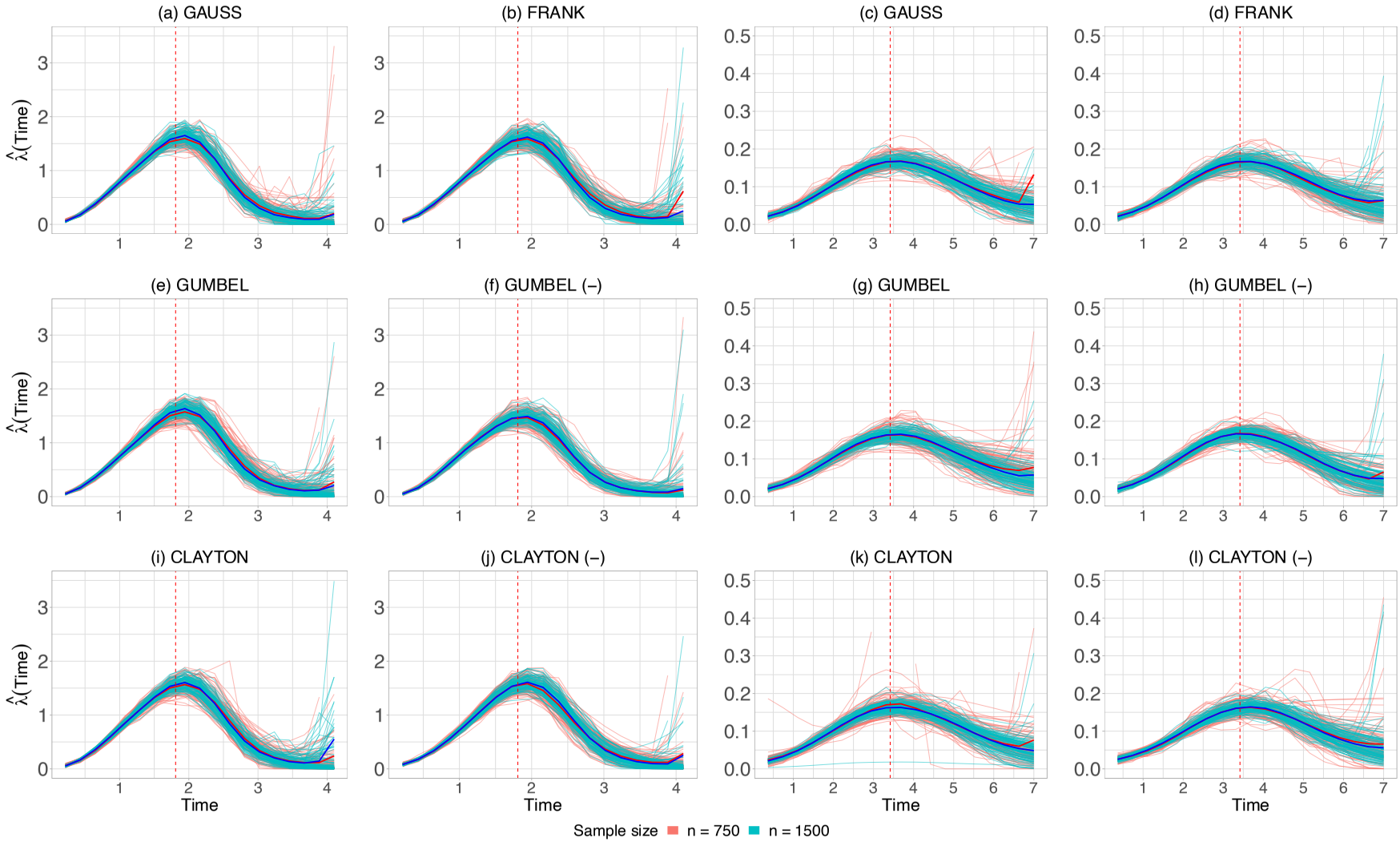

Figure F6: Estimated baseline hazard rate using the *PW* approach across sample sizes and copula functions given a discrete non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $t_{mild}^{max\ haz} = 1.809$ and $t_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

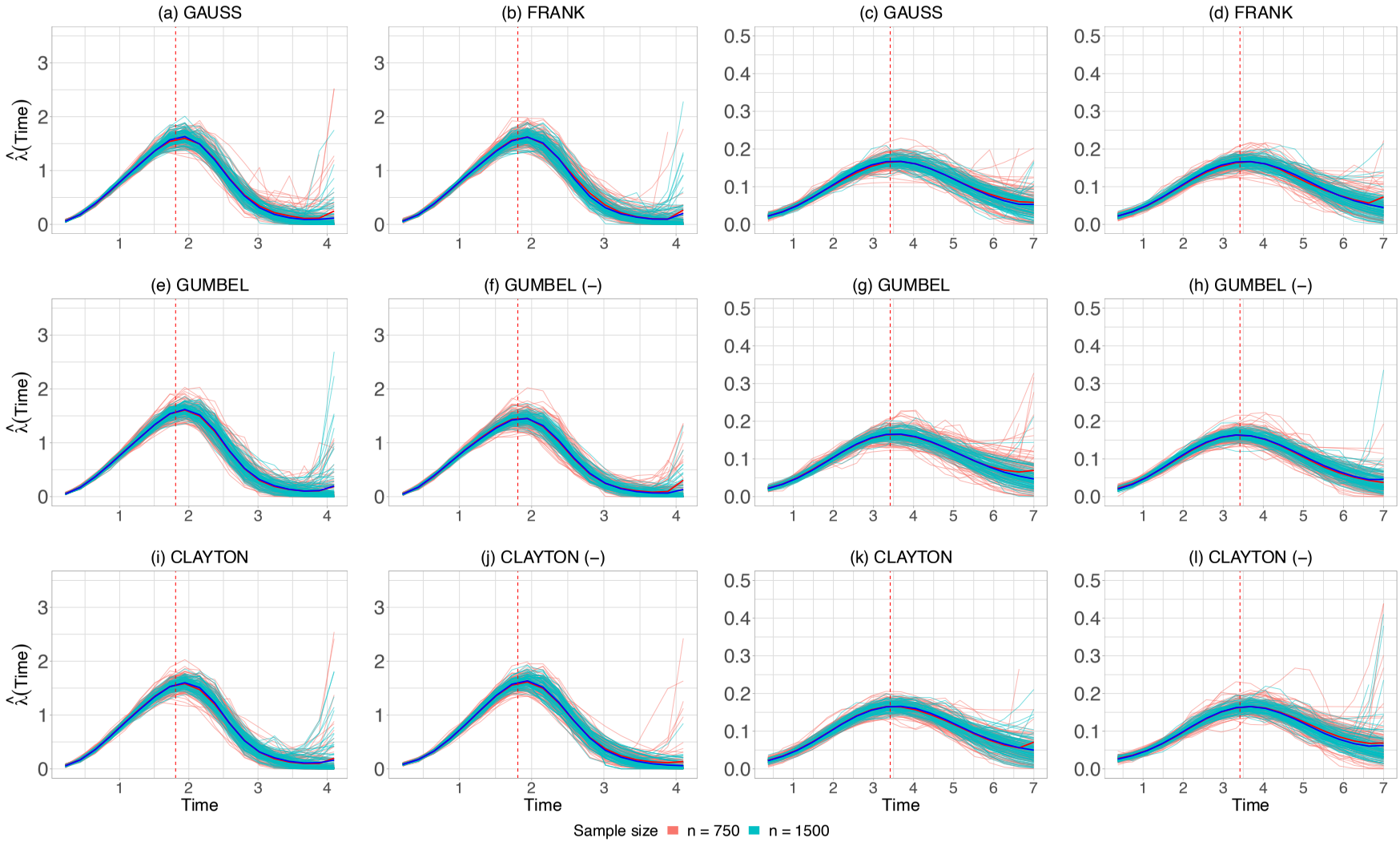

Figure F7: Estimated baseline hazard rate using the *PW* approach across sample sizes and copula functions given a continuous non-time-to-event margin. Solid blue and red lines indicate the average fit. Dotted red line indicates the time points at which the true hazard reaches its mode: $t_{mild}^{max\ haz} = 1.809$ and $t_{heavy}^{max\ haz} = 3.417$ for mild and heavy censoring rates. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

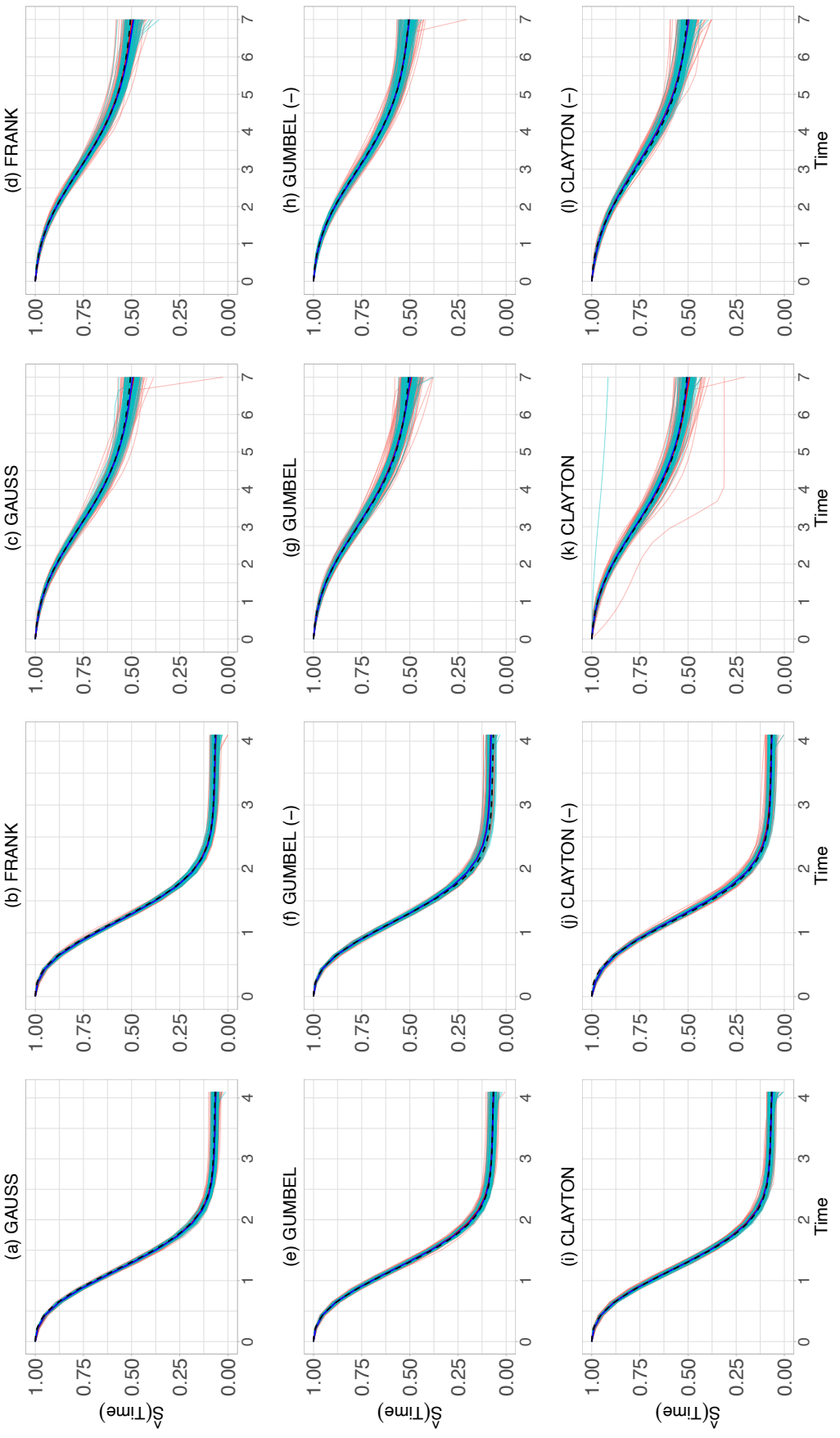

Figure F8: Estimated baseline survival function using the PW approach across sample sizes and copula functions given a discrete non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

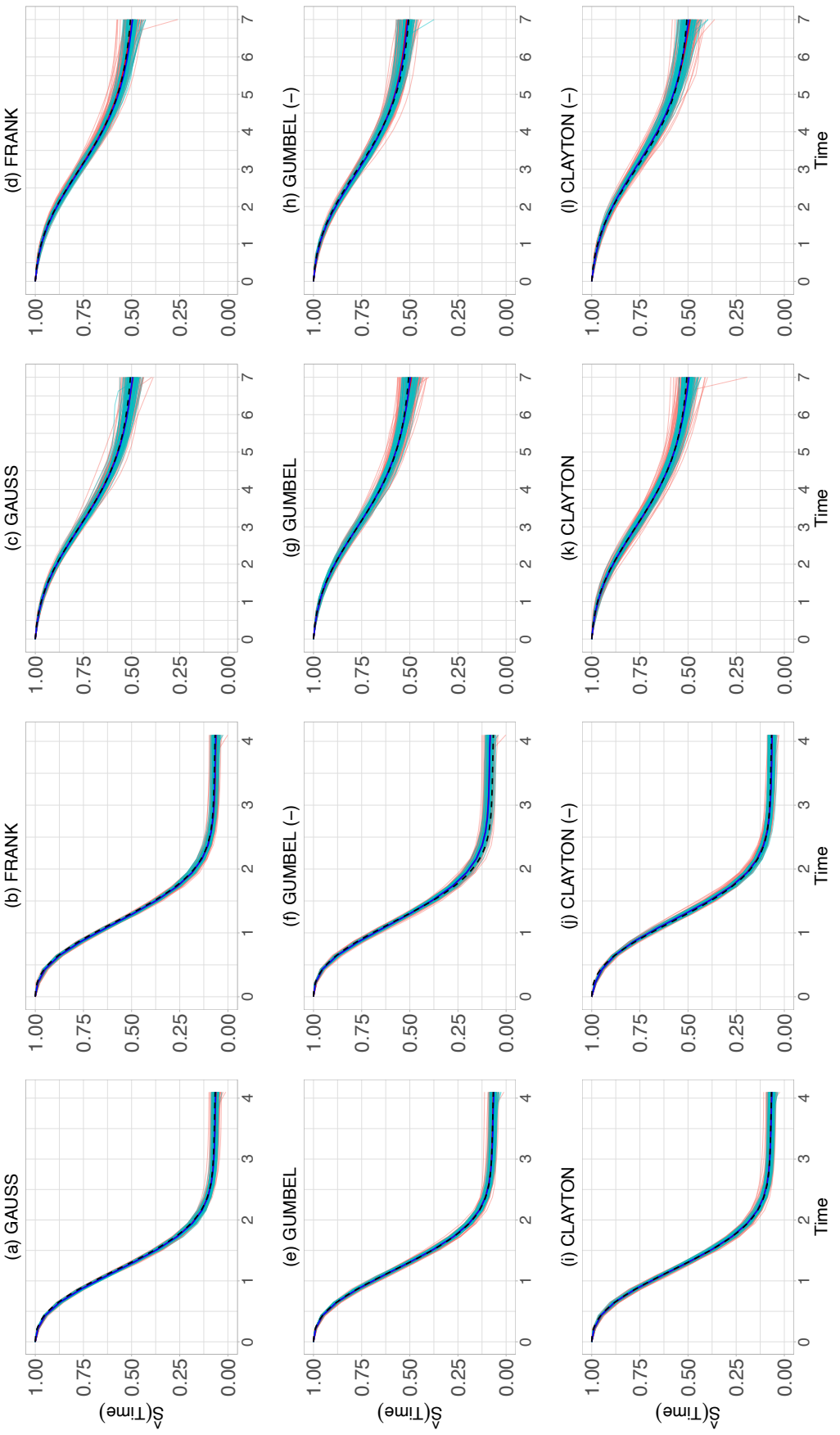

Figure F9: Estimated baseline survival function using the *PW* approach across sample sizes and copula functions given a continuous non-time-to-event margin. Solid blue and red lines indicate the average fit, black solid line indicates the true survival function. Mild censoring: (a), (b), (e), (f), (i), (j). Heavy censoring: (c), (d), (g), (h), (k), (l).

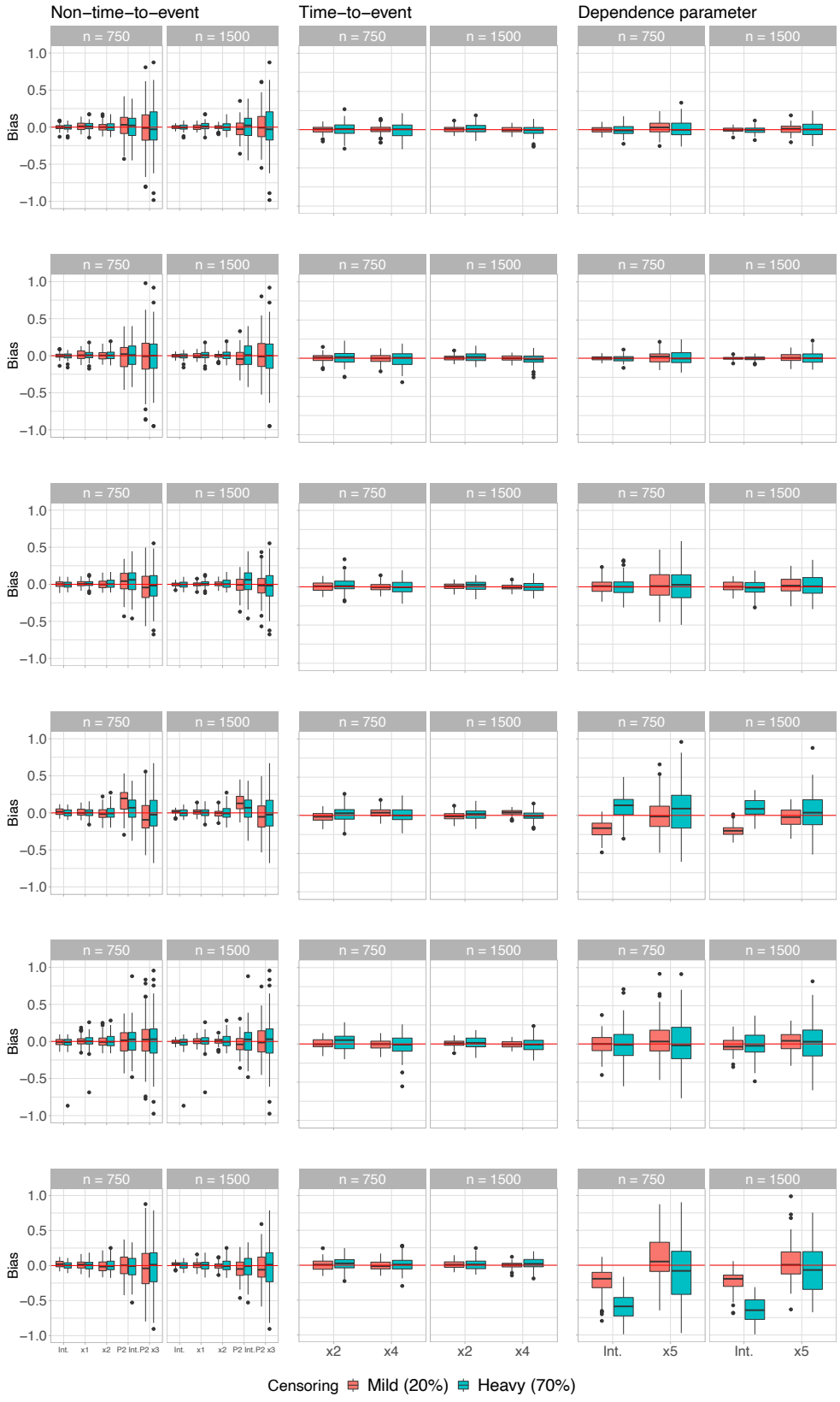

Figure F10: Bias of coefficients for a discrete non-time-to-event margin (DT). Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.

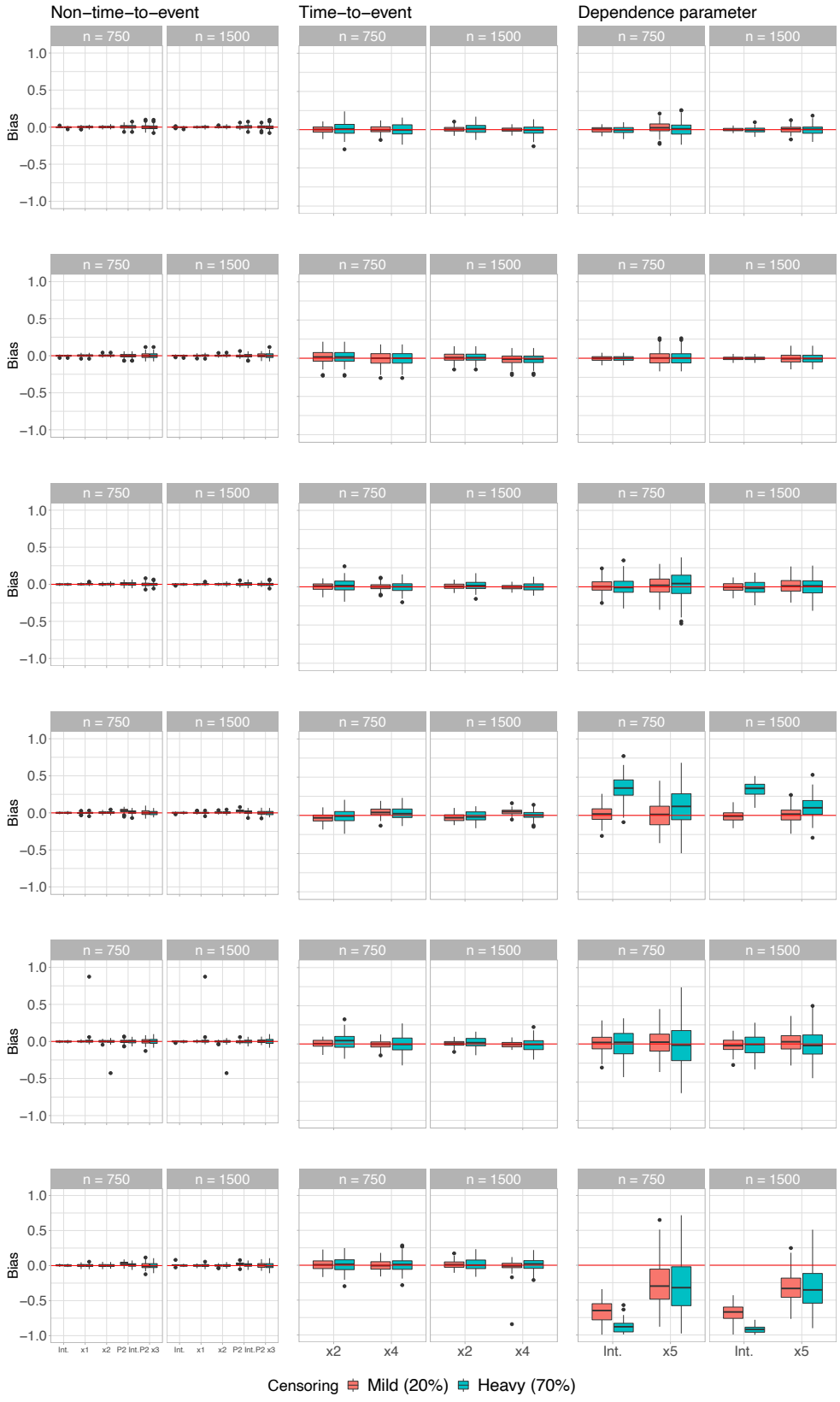

Figure F11: Bias of coefficients for a continuous non-time-to-event margin (DT). Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.

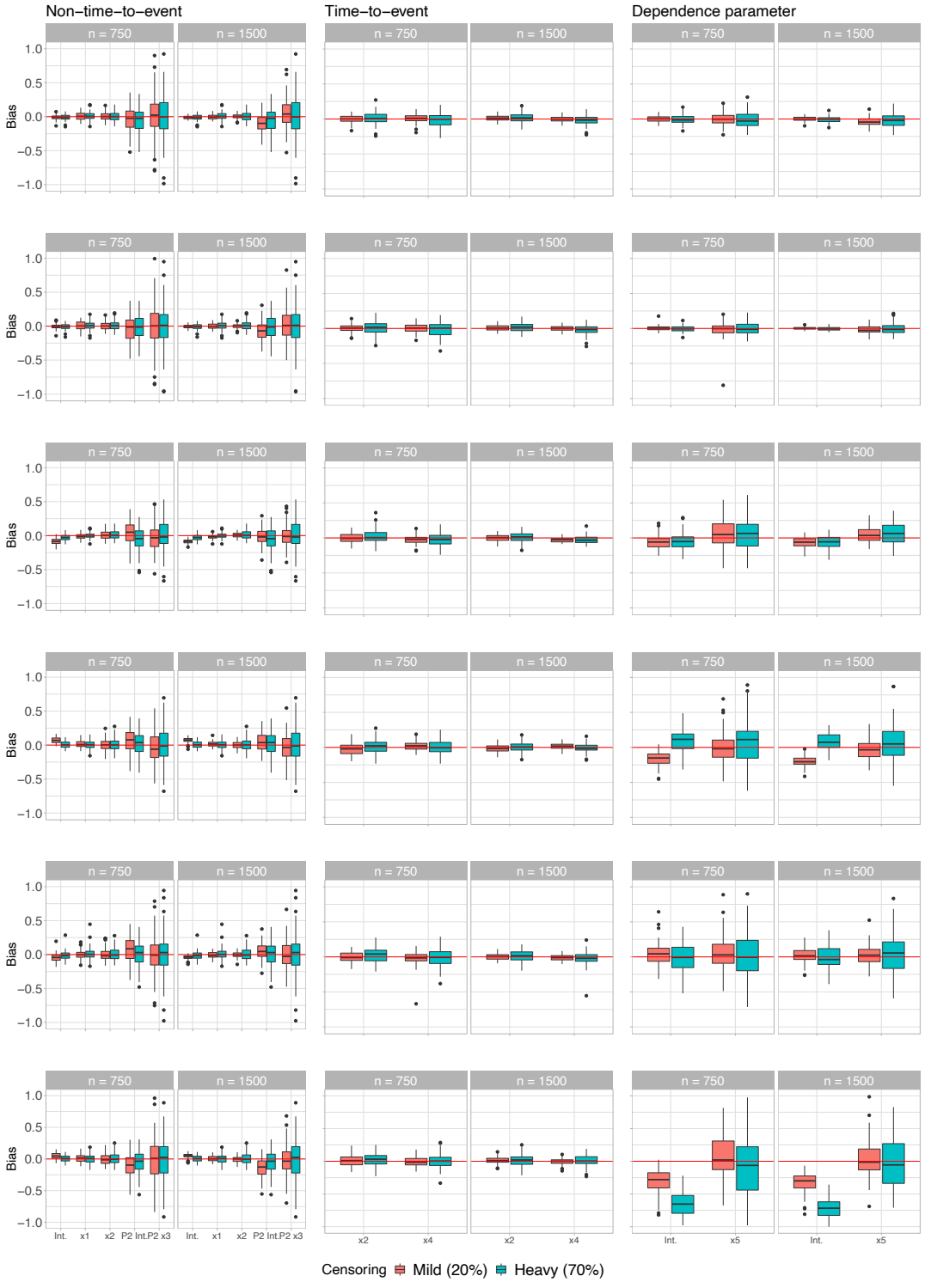

Figure F12: Bias of coefficients for a discrete non-time-to-event margin (PW). Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.

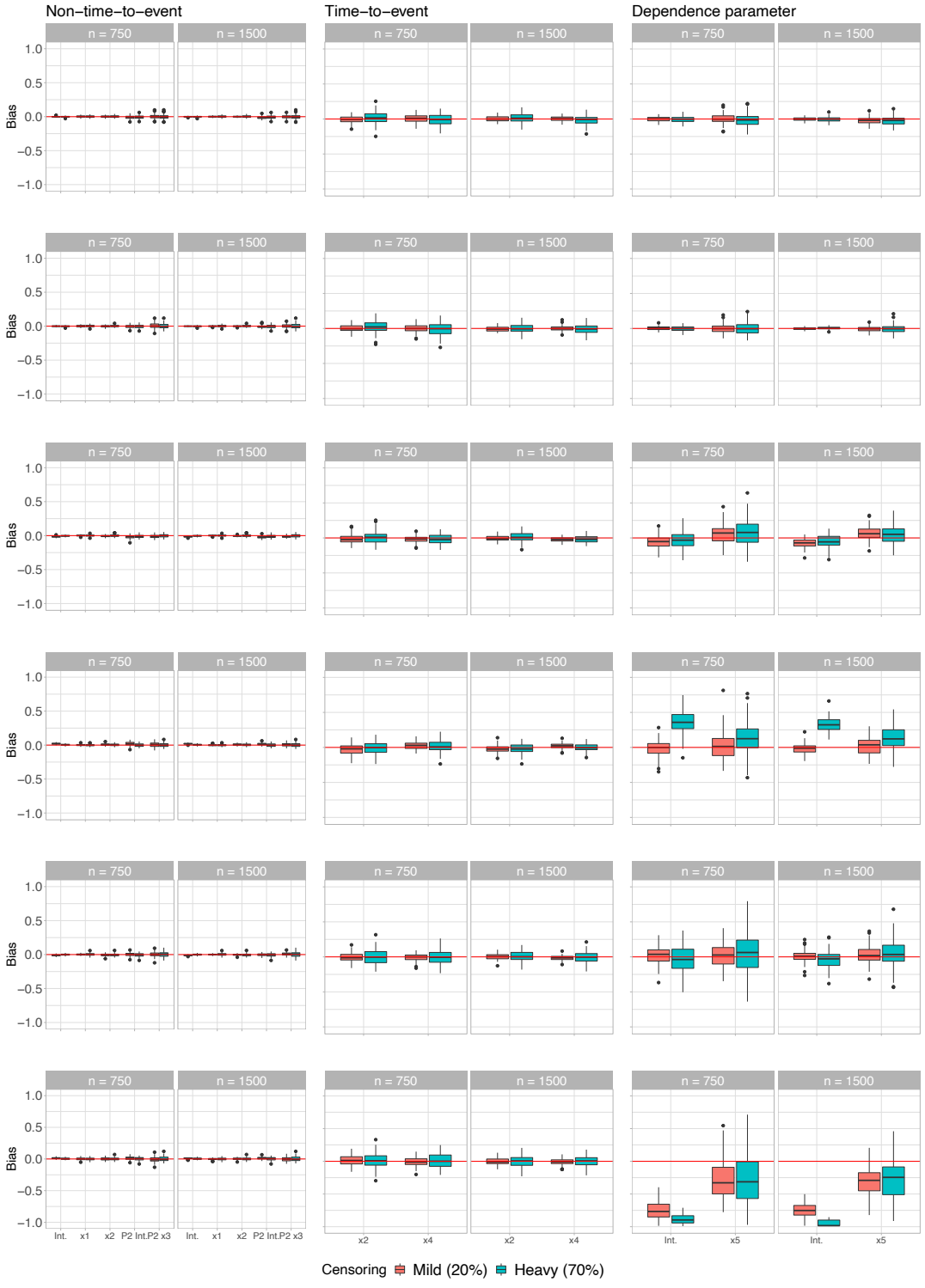

Figure F13: Bias of coefficients for a continuous non-time-to-event margin (PW). Rows indicate Gauss, Frank, Gumbel $(+,-)$, and Clayton $(+,-)$ copulas.