Received: 21 March 2022 | Accepted: 15 September 2022

DOI: 10.1111/biom.13765

Biometrics
A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY
WILEY

BIOMETRIC METHODOLOGY

Boosting distributional copula regression

Nicolai Hans[1] | **Nadja Klein**[1] | **Florian Faschingbauer**[2] | **Michael Schneider**[2] | **Andreas Mayr**[3]

[1]Chair of Statistics and Data Science, Humboldt-Universität zu Berlin, Berlin, Germany

[2]Department of Obstetrics and Gynecology, University Hospital of Erlangen, Erlangen, Germany

[3]Department of Medical Biometrics, Informatics and Epidemiology, Faculty of Medicine, University of Bonn, Bonn, Germany

Correspondence
Nadja Klein, Chair of Statistics and Data Science, Humboldt-Universität zu Berlin, Germany.
Email: nadja.klein@hu-berlin.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: KL3037/2-1, MA7304/1-1

Abstract

Capturing complex dependence structures between outcome variables (e.g., study endpoints) is of high relevance in contemporary biomedical data problems and medical research. Distributional copula regression provides a flexible tool to model the joint distribution of multiple outcome variables by disentangling the marginal response distributions and their dependence structure. In a regression setup, each parameter of the copula model, that is, the marginal distribution parameters and the copula dependence parameters, can be related to covariates via structured additive predictors. We propose a framework to fit distributional copula regression via model-based boosting, which is a modern estimation technique that incorporates useful features like an intrinsic variable selection mechanism, parameter shrinkage and the capability to fit regression models in high-dimensional data setting, that is, situations with more covariates than observations. Thus, model-based boosting does not only complement existing Bayesian and maximum-likelihood based estimation frameworks for this model class but rather enables unique intrinsic mechanisms that can be helpful in many applied problems. The performance of our boosting algorithm for copula regression models with continuous margins is evaluated in simulation studies that cover low- and high-dimensional data settings and situations with and without dependence between the responses. Moreover, distributional copula boosting is used to jointly analyze and predict the length and the weight of newborns conditional on sonographic measurements of the fetus before delivery together with other clinical variables.

KEYWORDS

Archimedean copula, component-wise gradient boosting, early stopping, GAMLSS, tail dependence

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

1 | INTRODUCTION

The analysis of complex association structures between multiple outcome variables is of increasing interest in contemporary biomedical research. For instance, in genetic epidemiology the joint consideration of multiple phenotypes leads to a better understanding of physical and mental disorders and the identification of relevant genetic

This is an open access article under the terms of the Creative Commons Attribution-NonCommercial-NoDerivs License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2021 The Authors. *Biometrics* published by Wiley Periodicals LLC on behalf of International Biometric Society.

risk factors (Ghosh, 2014; Ott & Wang, 2011). Moreover, in clinical medicine multivariate investigations provide a broader view on diseases like diabetes (Espasandín-Domínguez et al., 2019) or Alzheimer's disease (Yan et al., 2015) and important clinical measures like the body fat percentage (Petterle et al., 2021). The motivating example of this paper is to model and predict the height and weight of newborn babies based on sonographic and clinical covariates of fetuses and mothers collected at the Erlangen University Hospital before birth (Faschingbauer et al., 2016). In clinical obstetrics and gynecology, the prediction of the fetal weight is of high relevance for decision making during the birth process. We investigate the fetal weight and height together, as these measures are likely to interrelate. From a delivery management perspective the analysis of both responses yields new insights like the identification of cases with disproportional growth or the calculation of joint probabilities that crucial thresholds are passed. Moreover, it could also be of particular interest to explore predictor variables that influence the association between the weight and the length of the fetus.

To define an appropriate and flexible joint model, we employ copulas and integrate their use into a bivariate distributional regression model. Copula models offer a flexible approach to define multivariate distributions of several outcome variables. In particular, by using copulas the model-building process is conveniently decomposed into the specification of the univariate marginal distributions (i.e., in our case those of fetal weight and height) and the selection of an appropriate copula function that defines the dependence structure (Nelsen, 2006). Thus, response variables with potentially different marginal distributions can be combined via copula functions that introduce distinct dependence scenarios. This makes the copula approach a versatile tool for multivariate analysis going well beyond Gaussian distribution assumptions of the marginals or linear correlations. Within the structured additive distributional regression framework like generalized additive models for location scale and shape (GAMLSS; Rigby & Stasinopoulos, 2005), different types of univariate response variables, that is, continuous, discrete or mixed continuous-discrete, can be considered. This model class extends the generalized additive model (GAM; Hastie & Tibshirani, 1990) by associating every response distribution parameter with the covariates via additive predictors. The additive predictors allow for different covariate effects, that is, linear, nonlinear, random and spatial, on the model parameters (Wood, 2017). While there exists a rich literature regarding copula regression (see Craiu & Sabeti, 2012; Kolev & Paiva, 2009; Sabeti et al., 2014, and references therein), these typically cover only parts of the flexibility of the distributional copula regression frameworks introduced in a Bayesian (Klein & Kneib, 2016) and a penalized-likelihood (Marra & Radice, 2017) based context. As part of distributional copula regression all parameters of the model, that is, both the distributional parameters of the marginals and the dependence parameters, are related to the covariates via simultaneously estimated additive predictors. Hence, the dependence structure is also described by flexible covariate effects that possibly exceed classical linear relations.

Our contribution is to develop a general boosting approach to structured additive distributional copula regression. We refer to this framework as *boosted copula regression* and we conceptualize and implement a first scenario with continuous marginals and three different copula functions. To do so, we extend a popular estimation technique of modern data analysis called model-based boosting to fit distributional copula regression models. While the original concept of boosting arose in the machine learning community (Freund & Schapire, 1996), Friedman et al. (2000) provided a statistical view on boosting. Model-based boosting (Bühlmann & Yu, 2003; Bühlmann & Hothorn, 2007) builds on a functional gradient boosting paradigm (Friedman, 2001) and serves well for estimation problems in the context of univariate regression models. Mayr et al. (2012) further extended the model-based boosting algorithm to GAMLSS. The estimation via model-based boosting yields several advantages compared to the Bayesian and frequentist counterparts that are of high interest in applied data analysis. First, model-based boosting incorporates an intrinsic mechanism for variable selection. Variable selection, that is, the problem to select a small set of truly informative covariates, presents a key issue in applied statistics (Hastie et al., 2009). This becomes even more urgent in the context of complex model classes containing various additive predictors. The model-based boosting algorithm thereby automatically leads to sparse solution of the parameter-specific structured additive predictors. Second, fitting regression models via boosting shrinks the effect estimates toward zero. This leads to a decreased variability of predictor effects and typically also to an improvement in the prediction accuracy (Mayr et al., 2012). Finally and third, model-based boosting is capable to estimate regression models in high-dimensional settings, for instance, when the number of covariates exceeds the number of observations by far ($p \gg n$), as it frequently occurs in biomedical research (Bermingham et al., 2015; Romero et al., 2006). As a result, model-based boosting complements existing Bayesian and penalized-likelihood techniques in the context of copula regression models and incorporates valuable modeling features that are of high relevance in applied data analysis.

The paper is structured as follows. We first review distributional copula regression models with continuous margins in Section 2. In Section 3, we extend the

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

model-based boosting algorithm to the distributional copula regression class and give insights into the algorithmic structure and hyperparameter tuning. Subsequently, Section 4 summarizes the main outcome of our simulation studies, whereas the Supporting Information (SI) covers the respective detailed documentations to evaluate the overall estimation results, the variable selection accuracy and the model building process in low- and high-dimensional settings and scenarios with dependent and independent responses. In Section 5, we apply boosted distributional copula regression to jointly analyze the birth length and weight of fetuses by means of a birth cohort dataset. Boosted copula distributional models are of particular interest in this application as the responses are defined on the positive real line and thus require other marginal distributions than the normal distribution. Moreover, the dataset consists of 36 covariates, which makes simultaneous variables selection for the different additive predictors of the model parameters a non-trivial task that can be tackled with boosting. Finally, parameter shrinkage induced by boosting in particular suits prediction setups like in delivery management scenarios. We conclude with a thorough discussion and further research ideas in Section 6.

2 | DISTRIBUTIONAL COPULA REGRESSION MODELS

We now review structured additive distributional copula regression models along the lines of Klein and Kneib (2016). As these authors, we focus on the bivariate case throughout but will discuss possible extensions to multivariate situations at the end of the paper.

2.1 | The notion of distributional copula regression models

Copulas provide a flexible approach to the construction of multivariate distributions and regression models. One can show that the joint conditional cumulative distribution function (CDF) of the two continuous random response variables $\boldsymbol{Y} = (Y_1, Y_2)^\top$ given a p-dimensional covariate vector $\boldsymbol{x} = (x_1, \dots, x_p)^\top$ can be expressed as

$$F_{\boldsymbol{Y}}(y_1, y_2 | \boldsymbol{\vartheta}) = C\left\{F_1\left(y_1 | \boldsymbol{\vartheta}^{(1)}\right), F_2\left(y_2 | \boldsymbol{\vartheta}^{(2)}\right) | \boldsymbol{\vartheta}^{(c)}\right\}, \qquad (1)$$

where $\boldsymbol{\vartheta} = \{(\boldsymbol{\vartheta}^{(1)})^\top, (\boldsymbol{\vartheta}^{(2)})^\top, (\boldsymbol{\vartheta}^{(c)})^\top\}^\top \in \mathbb{R}^K$, $K = K_1 + K_2 + K_c$, is the vector of model parameters, $F_1(y_1 | \boldsymbol{\vartheta}^{(1)})$ and $F_2(y_2 | \boldsymbol{\vartheta}^{(2)})$ are the marginal CDFs of the two response variables, respectively, and $C(\cdot, \cdot | \boldsymbol{\vartheta}^{(c)})$ is a uniquely defined copula function (Patton, 2006; Sklar, 1959). The vectors $\boldsymbol{\vartheta}^{(\bullet)}$ with $\bullet \in \{1, 2, c\}$ contain the $k = 1, \dots, K_\bullet$ parameters $\vartheta_k^{(\bullet)}$ of the marginal distributions and the copula function. In our distributional regression setup, all components of $\boldsymbol{\vartheta}$ are linked to (potentially different subsets of) the covariate vector $\boldsymbol{x}$ via additive predictors and appropriate link functions. Since $F_1(y_1 | \boldsymbol{\vartheta}^{(1)})$ and $F_2(y_2 | \boldsymbol{\vartheta}^{(2)})$ can be interpreted as uniformly distributed random variables u_1 and u_2, respectively, $C(\cdot, \cdot | \boldsymbol{\vartheta}^{(c)})$ is a bivariate CDF on $[0, 1]^2$ that does not depend on the specific marginal CDFs. Further, if $C(\cdot, \cdot | \boldsymbol{\vartheta}^{(c)})$ is a conditional copula function and $F_1(y_1 | \boldsymbol{\vartheta}^{(1)})$ and $F_2(y_2 | \boldsymbol{\vartheta}^{(2)})$ are conditional CDFs, then $F_{\boldsymbol{Y}}$ is a conditional bivariate CDF with conditional marginal distributions F_1 and F_2. Hence, linking together two (different) univariate continuous distributions via a copula yields a valid bivariate distribution with conditional probability density function (PDF)

$$f_{\boldsymbol{Y}}(y_1, y_2 | \boldsymbol{\vartheta}) = c\left\{F_1\left(y_1 | \boldsymbol{\vartheta}^{(1)}\right), F_2\left(y_2 | \boldsymbol{\vartheta}^{(2)}\right) \mid \boldsymbol{\vartheta}^{(c)}\right\} \times f_1\left(y_1 | \boldsymbol{\vartheta}^{(1)}\right) f_2\left(y_2 | \boldsymbol{\vartheta}^{(2)}\right), \qquad (2)$$

where $f_1(y_1 | \boldsymbol{\vartheta}^{(1)})$ and $f_2(y_2 | \boldsymbol{\vartheta}^{(2)})$ are the marginal PDFs and

$$c\left\{F_1\left(y_1 | \boldsymbol{\vartheta}^{(1)}\right), F_2\left(y_2 | \boldsymbol{\vartheta}^{(2)}\right) | \boldsymbol{\vartheta}^{(c)}\right\} = \frac{\partial^2}{\partial F_1 \partial F_2} F_{1,2}(y_1, y_2 | \boldsymbol{\vartheta})$$

is the copula density. For n bivariate observations $\{(y_{i1}, y_{i2})^\top\}_{i=1,\dots,n}$ with covariate vectors $\boldsymbol{x}_1, \dots, \boldsymbol{x}_n$ and observation-specific distributional parameters $\boldsymbol{\vartheta}_{(i)} = \{(\boldsymbol{\vartheta}_{(i)}^{(1)})^\top, (\boldsymbol{\vartheta}_{(i)}^{(2)})^\top, (\boldsymbol{\vartheta}_{(i)}^{(c)})^\top\}^\top \in \mathbb{R}^K$, Equation (2) induces the joint log-likelihood function

$$\begin{aligned} l(\boldsymbol{\vartheta}_{(1:n)}) \equiv & \sum_{i=1}^{n} \log\left[c\left\{F_1\left(y_{1i} \mid \boldsymbol{\vartheta}_{(i)}^{(1)}\right), F_2\left(y_{2i} \mid \boldsymbol{\vartheta}_{(i)}^{(2)}\right) \mid \boldsymbol{\vartheta}_{(i)}^{(c)}\right\}\right] \\ & + \sum_{i=1; d \in \{1,2\}}^{n} \log\left\{f_d\left(y_{di} \mid \boldsymbol{\vartheta}_{(i)}^{(d)}\right)\right\}, \ \boldsymbol{\vartheta}_{(1:n)} \\ & = \left(\boldsymbol{\vartheta}_{(1)}^\top, \dots, \boldsymbol{\vartheta}_{(n)}^\top\right)^\top \in \mathbb{R}^{n \times (K)}. \end{aligned} \qquad (3)$$

2.2 | Marginal distributions

A variety of continuous marginals can be applied within our copula framework to account for, for example, skewness, heteroscedasticity, or heavy tails. In the empirical evaluation of Sections 4 and 5, the log-normal and the log-logistic distributions will be of relevance, which both comprise two distributional parameters. Further

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

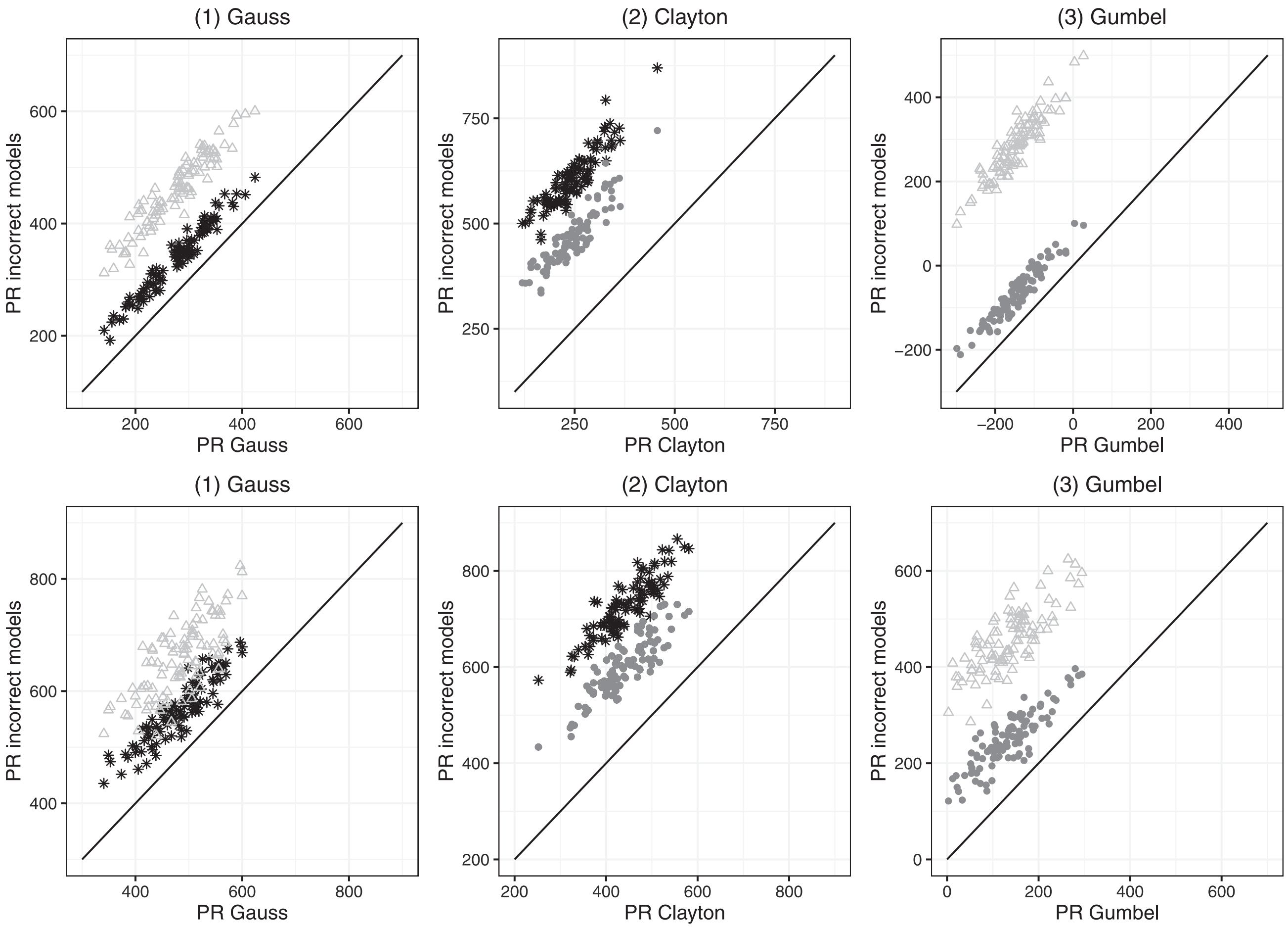

FIGURE 1 Predictive risk (PR) values of the correct model (*x*-axis) vs. the incorrect models (*y*-axis). For (1)–(3), the Gaussian, the Clayton, and the Gumbel copula are the correct model, respectively. The triangles (light grey) mark the comparison to the incorrect Clayton copula, the stars (black) to the incorrect Gumbel copula, and the points (dark grey) to the incorrect Gaussian copula. The plots of the low- and high-dimensional data setup are represented in the first and the second row, respectively.

information on the PDF, the CDF and the mean of both distributions are given in SI A.

2.3 | Dependence structure

Archimedean and elliptical copulas are popular copula function choices in practice (McNeil et al., 2005) and allow for different dependence scenarios (see SI B Web Figure 1 for a graphical illustration). To summarize the associations between the components of a multivariate response beyond the commonly employed linear Pearson's correlation coefficient, important dependence measures in a copula context are given by the Kendall's tau rank correlation τ^{κ} and the upper and the lower tail dependence, which are defined as $\lambda_u = \lim_{q\to 1^-} P\{y_2 > F_2^{-1}(q) | y_1 > F_1^{-1}(q)\}$ and $\lambda_l = \lim_{q\to 0^+} P\{y_2 \leq F_2^{-1}(q) | y_1 \leq F_1^{-1}(q)\}$, respectively (McNeil et al., 2005). The most important example of elliptical copulas is the Gaussian copula (Song, 2000), which has no tail dependence, that is, y_1 and y_2 are independent in the limit and $\lambda_u = \lambda_l = 0$. In contrast, Archimedean copulas do allow for lower or upper tail dependence. The Clayton and Gumbel copulas are prominent members of this family, the former allowing for lower tail dependence and the latter for upper tail dependence. Note that all introduced copula functions contain only one parameter of dependence, which we denote as $\vartheta^{(c)}$. An overview of the introduced copulas and the respective formulas for Kendall's tau and the lower and upper tail dependence are given in Table 1.

2.4 | Structured additive predictors

In distributional copula regression, each distributional parameter $\vartheta_k^{(\bullet)}$ can be associated with a structured additive predictor $\eta_k^{(\bullet)}$ through appropriate monotonic

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

TABLE 1 Details on the Gaussian, Clayton, and Gumbel copulas. The functions $\Phi(\cdot)$ and $\Phi_2(\cdot,\cdot;\rho)$ denote the CDFs of the standard univariate normal distribution and bivariate normal distribution with standard normally distributed margins and linear correlation parameter $\rho \equiv \vartheta^{(c)}$, respectively.

Copula	$\boldsymbol{C(u_1, u_2; \vartheta^{(c)})}$	**Tail dependence**	τ^{κ}
Gaussian	$\Phi_2\{\Phi^{-1}(u_1), \Phi^{-1}(u_2); \vartheta^{(c)}\}$	$\lambda_u = \lambda_l = 0$	$\frac{2}{\pi}\arcsin(\vartheta^{(c)})$
Clayton	$(u_1^{-1} + u_2^{-1} - 1)^{-1/\vartheta^{(c)}}$	$\lambda_u = 0,\ \lambda_l = 2^{\frac{-1}{\vartheta^{(c)}}}$	$\frac{\vartheta^{(c)}}{(\vartheta^{(c)}+2)}$
Gumbel	$\exp[-\{(-\log(u_1))^{\vartheta^{(c)}} + (-\log(u_2))^{\vartheta^{(c)}}\}^{1/\vartheta^{(c)}}]$	$\lambda_u = 2 - 2^{\frac{1}{\vartheta^{(c)}}},\ \lambda_u = 0$	$1 - \frac{1}{\vartheta^{(c)}}$

response functions $h_k^{(\bullet)}(\cdot)$ to ensure the restrictions on the respective parameter spaces. This leads to $\vartheta_k^{(\bullet)} = h_k^{(\bullet)}(\eta_k^{(\bullet)})$ and $\eta_k^{(\bullet)} = (h_k^{(\bullet)})^{-1}(\vartheta_k^{(\bullet)})$, where the inverse $(h_k^{(\bullet)})^{-1}$ is the link function and is similar to GAMs, where link functions are used to model the conditional expectation of a distribution from the exponential family. Now, in the spirit of GAMLSS, each structured additive predictor is modeled through a sum of an overall intercept term $\beta_{0,k}^{(\bullet)}$ plus $J_k^{(\bullet)}$ generic functions $f_{j,k}^{(\bullet)}(\boldsymbol{x}_j)$ expressed as

$$\eta_k^{(\bullet)} = \beta_{0,k}^{(\bullet)} + \sum_{j=1}^{J_k^{(\bullet)}} f_{j,k}^{(\bullet)}(\boldsymbol{x}_j),$$

where each parameter specific predictor is associated to an individual subset of covariates, that is, $\boldsymbol{x}_1, \dots, \boldsymbol{x}_{J_k^{(\bullet)}}$. The intercept is the overall level of the predictor when all function evaluations are zero and the effects of the covariates on the parameter models are determined by the functions $f_{j,k}^{(\bullet)}(\boldsymbol{x}_j)$. For instance, in the application in Section 5 we fit models that incorporate linear and nonlinear effects. Dropping the parameter superscripts $(\bullet)$ and parameter index k a single generic linear effect is given by a function of the form $f_j^{\text{linear}}(x_j) = \beta_j x_j$. Nonlinear effects of univariate covariates x_j can be modeled via P-splines with L_j B-Spline basis functions $B_{j,l_j}(x_j)$ resulting in the generic smooth function $f_j^{\text{smooth}}(x_j) = \sum_{l_j=1}^{L_j} \beta_{j,l_j} B_{j,l_j}(x_j)$. A second order difference penalty on the coefficients is introduced to control for the smoothness of the nonlinear effect (Eilers & Marx, 1996). Many types of smoothing functions, such as smoothing or regression splines, exist (Wood, 2017). Moreover, further covariate effect types like individual or group specific random effects or spatial effects can be incorporated in the structured additive predictors (Fahrmeir et al., 2013). In summary, each model parameter specific structured additive predictor is related to an individual set of covariates, which are also based on different functional forms.

3 | ESTIMATION VIA MODEL-BASED BOOSTING

Our proposed method includes the estimation of distributional copula regression models via model-based boosting. This section provides an overview on the algorithm and covers questions related to hyperparameter tuning and model building.

3.1 | Algorithm for boosting distributional copula models

Parameter estimation of the distributional copula regression model of a bivariate continuous random variable $\boldsymbol{Y}$ can be expressed in terms of the optimization problem

$$\hat{\boldsymbol{\eta}} = \underset{\boldsymbol{\eta}}{\arg\min} \left[E_{\boldsymbol{Y}} \left\{ \omega\left(\boldsymbol{Y}; \eta_1^{(1)}, \dots, \eta_{K_1}^{(1)}, \eta_1^{(2)}, \dots, \eta_{K_2}^{(2)}, \eta_1^{(c)}, \dots, \eta_{K_c}^{(c)} \right) \right\} \right],$$

where $\boldsymbol{\eta} = (\eta_1^{(1)}, \dots, \eta_{K_1}^{(1)}, \eta_1^{(2)}, \dots, \eta_{K_2}^{(2)}, \eta_1^{(c)}, \dots, \eta_{K_c}^{(c)})^\top \in \mathbb{R}^K$ is the vector of additive predictor functions of the model parameters, the respective estimates of the additive predictor functions are given by $\hat{\boldsymbol{\eta}} = (\hat{\eta}_1^{(1)}, \dots, \hat{\eta}_{K_1}^{(1)}, \hat{\eta}_1^{(2)}, \dots, \hat{\eta}_{K_2}^{(2)}, \hat{\eta}_1^{(c)}, \dots, \hat{\eta}_{K_c}^{(c)})^\top \in \mathbb{R}^K$ and $\omega(\cdot)$ denotes the loss function. Considering a data sample with observations $i = 1, \dots, n$, we minimize the empirical risk $\frac{1}{n}\sum_{i=1}^n \omega(\boldsymbol{y}_i; \eta_{(i),1}^{(1)}, \dots, \eta_{(i),K_1}^{(1)}, \eta_{(i),1}^{(2)}, \dots, \eta_{(i),K_2}^{(2)}, \eta_{(i),1}^{(c)}, \dots, \eta_{(i),K_c}^{(c)})$ over $\boldsymbol{\eta}_{(i)} = (\eta_{(i),1}^{(1)}, \dots, \eta_{(i),K_1}^{(1)}, \eta_{i1}^{(2)}, \dots, \eta_{(i),K_2}^{(2)}, \eta_{i1}^{(c)}, \dots, \eta_{(i),K_c}^{(c)})^\top \in \mathbb{R}^K$ instead. The loss function $\omega(\cdot)$ measures the discrepancy between the observed responses and the estimated additive predictors. A common choice for ω is the negative log-likelihood of the (bivariate) response distribution from Equation (3).

The fundamental idea of model-based gradient boosting is to sequentially minimize the empirical risk by a stepwise descent of the loss function's gradient in function space. It represents a sequential ensemble method, where the base-learners building the ensemble consist of regression functions. In more detail, the algorithm starts

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

with an initial specification of the additive predictors and a set of pre-defined regression-like base-learners, for example, simple linear models or regression splines with low degrees of freedom, for every model parameter. Note that the specification of the base-learner determines the type of effect the corresponding covariate has in the respective additive predictor (e.g., linear or nonlinear effect). An overview on potential base-learners in the context of model-based boosting is provided in Hofner et al. (2014). In every iteration m, the parameter-specific sets of base-learners are fitted one-by-one to the respective negative partial gradient vectors based on the fit from iteration $m-1$. These gradient vectors can be seen as pseudo-observations representing in the simplest case of the L_2 loss the residuals from the previous boosting iteration. In our case, they are the first derivatives of the negative log-likelihood with respect to the corresponding model parameter predictors $\eta_k^{(\bullet)}$. Subsequently, the best base-learner to fit the corresponding gradient vector in terms of a residual sum of squares criterion is selected for each parameter $\vartheta_k^{(\bullet)}$, which eventually leads to an intrinsic variable selection. Two algorithmic versions for boosting distributional regression models exist, namely the *cyclic* (Mayr et al., 2012) and the *noncyclic* (Thomas et al., 2018) algorithm. The *cyclical* version updates each additive predictor function with the respective best-fitting base-learner in each iteration successively, using the other estimates as offset values. The *noncyclic* version incorporates an additional selection step of the best-fitting model parameter. More specifically, in each iteration only the parameter model that yields the highest overall loss reduction is updated with its parameter-specific best base-learner. A strictly additive aggregation of the best fitting base-learners over the course of the boosting procedure accounts for the additive structure of the resulting predictor functions (see Section 2.4). For reasons of faster tuning properties, we focus on the *noncyclic* algorithm in the following. The algorithm is summarized in a generic way in SI C.

3.2 | Choice of hyperparameters

Two hyperparameters of the boosting algorithm are crucial for the parameter estimation and the complexity of the additive predictors, namely the step length s and the stopping iteration m_{stop}. The step length is involved in the additive updates, adding only a small proportion of the base-learner fit to the models. This guarantees the stability of the boosting algorithm and shrinks the parameter estimates to zero in combination with an appropriate stopping criterion. The stopping iteration m_{stop} defines the iteration after which further updates are no longer necessary. Early stopping, that is, when the algorithm is stopped before convergence, automatically controls the model complexity of the parameter specific additive predictors. Particularly, component-wise gradient boosting constitutes an intrinsic variable selection mechanism by assigning exactly one base-learner to each covariable $x_1, \dots, x_p$ for each model parameter (Bühlmann & Yu, 2003; Mayr et al., 2012). Consequently, in each iteration only the most informative variable is selected. Stopping the algorithm early finally leads to the inclusion of the most-informative variables in the additive predictors, while the less informative covariates are dismissed.

Tuning is the task of finding appropriate values for the hyperparameters. In the model-based boosting framework, the optimal m_{stop} is usually determined via cross-validation or resampling procedures based on the predictive risk (Bühlmann & Hothorn, 2007; Hothorn et al., 2005; Mayr et al., 2012). More precisely, the original data are split into a training and test fold several times. While a boosting model is fitted to the training sets, the test folds are used to assess the performance of the model by means of the predictive risk. Eventually, the optimal m_{stop} is assigned to the iteration of the smallest predictive risk averaged over the different test folds. In case of small datasets, it is also possible to subsequently leave out only one observation from fitting the model and perform this leave-one-out cross validation for all n observations. For large sample sizes also initially splitting the data into training and test data is an option. Note that a selection of m_{stop} based on information criteria like the Akaike information criterion (AIC) in contrast is problematic due to biased estimates for the effective degrees of freedom in case of boosting (Hastie, 2007; Mayr et al., 2012). The step length s might be set to a fix small value (much less than 1) in advance (Mayr et al., 2012). The simulation studies from Section 4 suggest that a step length of 0.01 is suitable for one parametric conditional copula models with two parametric marginal distributions.

3.3 | Model building

In the practical use of distributional copula regression, researchers are confronted with the choice of the marginal distributions and the copula function. Klein et al. (2019) argued that within the framework the model building process can be simplified by separating the task of finding the marginals and the copula function based on Bayesian information criteria in combination with normalized quantile residuals. Similarly, in a penalized likelihood-based context, the AIC and/or Bayesian information criterion (BIC) is proposed for those model building decisions (Marra and Radice, 2017). However, the boosting algorithm yields regularized fits, which makes the evaluation of residuals and information criteria like the AIC and BIC difficult

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

as mentioned above. As a consequence, for boosted univariate regression models authors suggest to consider the predictive risk, that is, the risk on a new dataset, to make decisions not only about m_{stop} but also the appropriate distribution function (Mayr et al., 2012). We follow their approach for the decision concerning the marginal distributions and expand the concept of the predictive risk also to the choice of the copula function. As a consequence, in our application in Section 5 we split the dataset randomly into training and test data. In a first step, we make a decision on the marginal distributions based on the predictive risk of the test data. Subsequently, we also choose the most appropriate copula function by means of the predictive risk using the same test data.

4 | SIMULATIONS

Estimating regression models via component-wise gradient boosting yields appealing properties like parameter shrinkage, an intrinsic variable selection mechanism and the capability to fit models in high-dimensional data. The following simulation study investigates these properties in the context of distributional copula regression with continuous margins. More specifically, we are interested in answering three questions, which are crucial for the application of boosted copula regression in practice:

1. Does the boosting algorithm estimate the correct (but shrunk) effects of the informative covariates on the different parameter models?
2. Is boosting able to separate the informative from the non-informative covariates?
3. Is the predictive risk a valid tool to choose the correct copula function?

We carry out simulations in `R` (R Core Team, 2020) that cover a low ($p = 20$) and a high-dimensional ($p = 1000$) data setup for the Gaussian, the Clayton and the Gumbel copula in combination with the log-normal and the log-logistic marginal distributions and $n = 1000$ observations. Note that the effective number of covariates is 100 and 5000 for the low- and the high-dimensional scenario, respectively, as each of the covariates might be integrated in each of the five sub-models of the copula distribution. We associate each model parameter with informative covariates in a linear and nonlinear fashion. However, most of the p covariates remain uninformative in both scenarios. For insights on the simulation setup and a more detailed discussion of the simulation results, we refer to SI D.2. Moreover, we confirm for a low-dimensional example with only four linear covariate effects that our algorithm converges to the estimates of the penalized likelihood approach of Marra and Radice (2017) (implemented in the R-package `GJRM`) for all three copulas and find a slightly better predictive performance of boosting once we apply early stopping (see SI D.1.1, D.1.2).

1. **Parameter estimates:** The boosting estimates reflect the true structure of the informative covariates on each parameter in both the low- and the high-dimensional settings, including a slight shrinkage toward zero. In general, the parameter shrinkage is more pronounced for the copula parameter in comparison to the marginal parameters and in the high-dimensional in comparison to the low-dimensional setting. A graphical representation of the estimates is provided via Web Figures 5–8 in the SI D.2.
2. **Variable selection:** The variable selection performance is sound for both the low- and the high-dimensional setting. In each simulation run, all informative variables were correctly selected for every model parameter, resulting in correct selection rates of 100%. However, in case of the low-dimensional setting the μ-models also tend to include many non-informative variables. In general, more uninformative variables are included in the low-dimensional setting compared to the high-dimensional scenarios. The selection rates of the non-informative covariates for the low- and high-dimensional setting are presented in Web Tables 2 and 3 in SI D.2, respectively.
3. **Predictive risk:** The predictive risk is a helpful tool to discriminate between the true and incorrect copula functions in the low- and the high-dimensional data case. Figure 1 displays the values of the predictive risk for the correct copula specification on the x-axis versus the values of the predictive risk of the misspecified models on the y-axis for each copula function. All points for all three correct copulas are located above the diagonal line, indicating that the true model is always selected for both the low- and the high-dimensional setting. Note that for a correct selection the predictive risk of the true model needs to be smaller than the predictive risk value of the misspecified models.

Finally, we consider a more extreme high-dimensional scenario ($p = 1000$), with very little observations ($n = 200$). As expected, the results for estimation, variable selection, and predictive risk get worse when n gets too small (see SI D.2.4) and the algorithm tends to select too simplistic models to avoid overfitting.

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

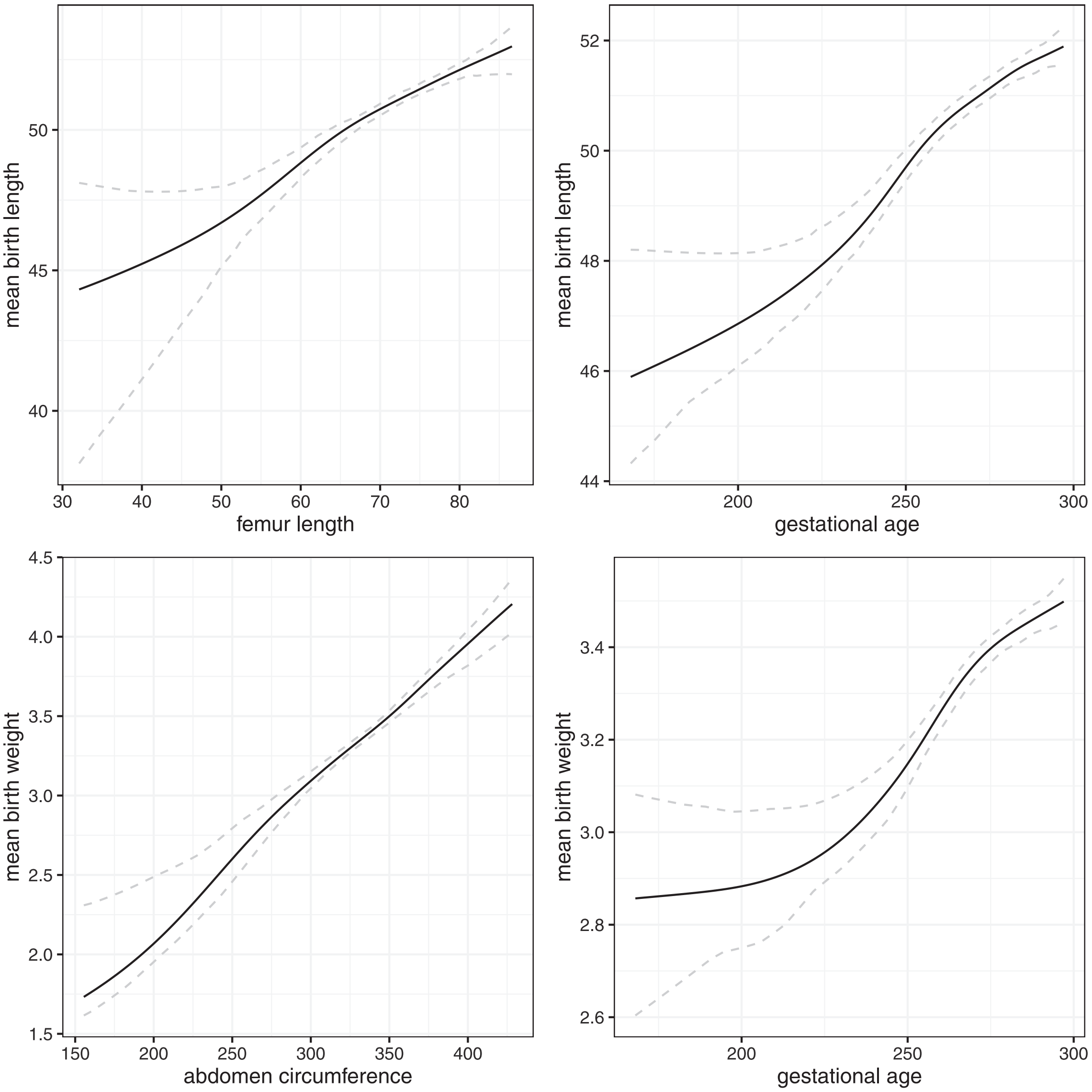

FIGURE 2 Effects of femur length and gestational age on the mean birth length in the first row of plots and effects of abdomen circumference and gestational age on the mean birth weight in the second row of plots (solid line). The dashed lines represent the 95% confidence bands estimated from 100 bootstrap samples (Hofner et al., 2016).

5 | ANALYSIS OF FETAL ULTRASOUND DATA

In this section, we model and predict the height and weight of newborn babies based on sonographic and clinical covariates of fetuses and mothers collected at the Erlangen University Hospital before birth (Faschingbauer et al., 2016) by means of boosted distributional copula regression. The weight of a fetus is an important factor in clinical obstetrics and gynecology (Barker, 1997). Both a very low and a very high weight are associated with increased risks of adverse events during labor. Fetal ultrasound measurements play an important role in the determination of the fetal weight and thus are of major importance for delivery and labor management (Dudley, 2005). In the current literature and clinical practice, the length of the newborn is not part of the prediction scheme—but could also lead to a better decision making during the labor process. Model-based boosting was applied several times in the development of prediction formulas for fetal weight, due to its advantages in prediction situations and its intrinsic variable selection mechanism (Faschingbauer et al., 2012, 2015; Schild et al., 2008), but we are the first to consider fetal weight jointly with its height.

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

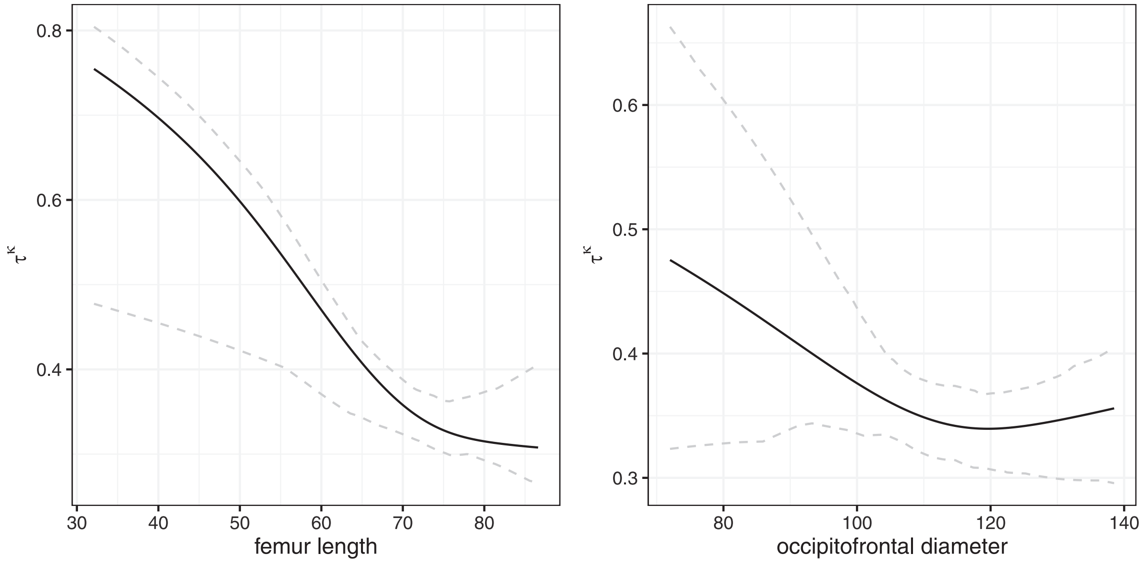

FIGURE 3 Effects of femur length and occipitofrontal diameter on Kendall's tau rank correlation coefficient τ^{κ} between birth length and weight (solid line). The dashed lines represent the 95% confidence bands estimated from 100 bootstrap samples (Hofner et al., 2016).

5.1 | Birth cohort data

Previous analyses investigate different facets of sonographic fetal weight prediction by means of the Erlangen birth cohort data (Faschingbauer et al., 2015, 2016). The dataset we use for our illustrative analysis was collected at the Department of Obstetrics and Gynecology of the Erlangen University Hospital and consists of $n = 6103$ pregnancy observations during 2008–2016. We include $p = 36$ covariates consisting of seven sonographic variables and their interaction terms as well as various clinical variables of the mother and the fetus. The response variables birth length and birth weight are measured in centimeters and kilograms, respectively. For a detailed description of the dataset and the covariates, we refer to SI E.1.

5.2 | Model building

First of all, we are concerned with the choice of the marginal distributions and the copula function. Following Section 3.3, we randomly assign the data to a training ($n_{\text{train}} = 4103$) and a test ($n_{\text{test}} = 2000$) dataset and choose the optimal marginal distributions and the copula function by means of the predictive risk. For both response variables, we compare the performance of the Gamma, inverse Gaussian, log-normal, log-logistic, and Weibull distribution, as birth length and birth weight are continuous on $\mathbb{R}^{+}$. To ensure similar effective step-lengths among outcomes and parameter dimensions, we apply a gradient stabilization (Hofner et al., 2016). Regarding the type of effects, in all model fits we use cubic P-splines with 20 equidistant knots, a second-order difference penalty and four degrees of freedom as base-learners for all continuous variables (allowing for nonlinear effects on the corresponding model parameter) and include the categorical variables' sex of the fetus and gestational diabetes via linear base-learners. For both birth length and birth weight, the log-logistic distribution suits best (see SI E.2 Web Table 6). Next, we decide on the most appropriate copula function by fitting the Gaussian, the Clayton, and the Gumbel copula with the selected log-logistic marginal distributions to the data. The Gaussian copula is the most appropriate choice for this data situation (see SI E.2 Web Table 7).

5.3 | Results and selected variables

5.3.1 | Marginal distributions

For the predictors of the scale parameters μ of the log-logistic distributions for birth length and birth weight, the boosting algorithm selects 92% and 81% of the covariates, respectively. Less covariates are included into the predictors of the shape parameters σ with 44% for the birth length and 39% for the birth weight.

The plots of Figure 2 show the effects of femur length and gestational age on the mean birth length and abdominal circumference and gestational age on the mean of birth weight. For each plot, all other covariates are set to the reference category in case of the binary variables and set to the average covariate values of the whole dataset in case of continuous variables. The mean birth length and the mean

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

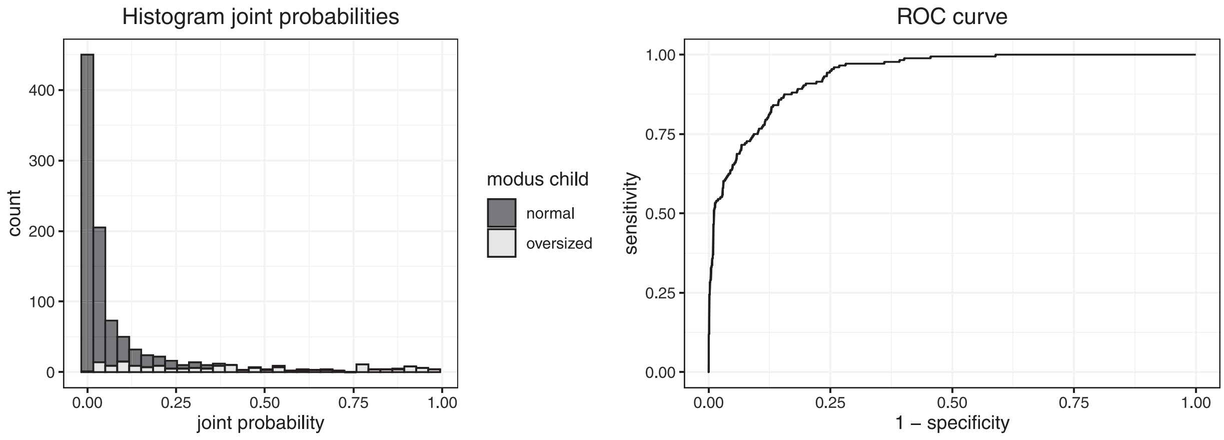

FIGURE 4 Joint probabilities from the copula model. Left: histogram of joint probabilities separated by oversized children (grey) and children that meet standard values for length and weight (black) as defined in the main text. Right: receiver operating characteristic (ROC) curve over the joint probabilities.

birth weight increase with higher covariate values for all presented plots. The effects of the sonographic covariates femur length and abdomen circumference are almost linear. These results are coherent with the findings of previous univariate sonographic birth weight estimations (Faschingbauer et al., 2012).

5.3.2 | Dependence parameter and Kendall's tau

The boosting algorithm selects 25% of the covariates into the dependence parameter model.

For both femur length and occipitofrontal diameter, Kendall's tau rank correlation coefficient τ^{κ} decreases for higher covariate values as presented in Figure 3. These findings seem plausible, as an increase of femur length possibly has a stronger impact on the birth length than on the birth weight. For the occipitofrontal diameter, we assume that a larger head of a fetus rather effects the birth weight than the birth length, leading to a weaker association.

5.4 | Comparison to univariate models

We want to compare the probabilistic forecasting performance of the Gaussian copula model with univariate distributional regression models for the birth length and the birth weight by fitting the two univariates' responses independently from each other. In order to compare the bivariate model with the univariate counterparts, we use multivariate proper scoring rules that assign a numerical score to a combination of a future observation and its predictive probabilistic distribution (Gneiting & Raftery, 2007). In particular, we focus on two scoring rules, namely the logarithmic score and the energy score, and calculate the averaged score values of the 2000 observations in the test data for the Gaussian copula model and the independent univariate models. Both scores favor the Gaussian copula model over the independent univariate models for our dataset. The resulting values for the logarithmic score are 2.079 and 2.228 for the Gaussian copula and independent models, respectively. For the energy score, the former has an average score of 1.085, while the latter yields an average score of 1.087. Note that for both scores lower values indicate a better probabilistic forecasting performance (Jordan et al., 2019).

5.5 | Joint probabilities and thresholding

Finally, we evaluate the predictive joint probabilities of the fetuses from the test sample to exceed a birth length of 51 cm and a birth weight of 4 kg for the Gaussian copula model. A birth weight of >4 kg is defined as fetal macrosomia (Faschingbauer et al., 2015). Following a study on international standards on newborns' length (Villar et al., 2014) the 90th centile of the birth length is given by approximately 51 cm for fetuses within the 38th and 41th week of gestational age. These two thresholds are interesting to investigate as it is for example more likely for mothers with large babies to experience a cesarean section (Boulet et al., 2003). The Erlangen birth dataset confirms this finding by a section rate that is approximately 1.42 times higher for newborns that exceed these thresholds versus newborns that meet standard values, that is, birth length $\in [47, 51]$ cm and birth weight $\in [2.65, 4]$ kg, where 47 cm and 2.65 kg are approximately the 10th centiles for birth length and birth weight for fetuses within the 38th and 41th

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

week of gestational age, respectively (Villar et al., 2014). Figure 4 (left) shows the histogram of the joint probabilities separated by oversized children (grey) and children that meet standard values for length and weight (black) in the test dataset. It is obvious that the joint probabilities to exceed both thresholds tend to be higher for oversized children. Thus, we consider the receiver operating characteristic curve (ROC, right) to evaluate the classification performance of the joint probabilities. With an area under the curve (AUC) of 0.937 the joint probability seems to be a valuable indicator to identify oversized children.

6 | CONCLUSION

We have introduced a flexible framework to fit conditional copula regression models via model-based boosting. Boosting proves to be a valuable fitting procedure in many applied data analysis settings, due to its fully data-driven mechanism to select variables and predictor effects, its feasibility in high-dimensional data settings and its good performance in prediction setups. In extensive simulation studies, we demonstrate the sound performance of boosted conditional copula regression models in low- and high-dimensional settings. Moreover, an application on fetal ultrasound data exemplary illustrates the potential of boosted copula regression in medical research. Boosted copula regression could be of interest in many other applications of biomedical research and beyond. In many clinical trials or experiments, one typically evaluates a potential intervention effect on various outcome variables (primary or secondary endpoints), which are typically analyzed separately. This may lead to insufficient results, since separate analyses cannot detect potential effects on the associations of these endpoints.

Despite its merits, we notice three major limitations of our approach that frequently occur in the context of boosting regression models in general. First of all, even though model-based boosting incorporates an intrinsic variable selection mechanism, in some situations the algorithm tends to include too many variables due to optimizing the predictive risk. This occurs in particular in low-dimensional data settings with large n, where boosting shows a slow overfitting behavior and a relatively late stopping of the algorithm (Staerk & Mayr, 2021). Strömer et al. (2022) investigated the latter phenomenon for univariate GAMLSS and proposed a new procedure for enhanced variable selection that effectively deselects base-learners with minor importance. This idea could be further developed for boosting distributional copula regression. Moreover, when boosting is applied to high-dimensional data with p larger than n, the algorithm tends to stop relatively early leading to sparse and thus simpler models—as can be also observed in our simulation results with smaller n. Note that for complex models like the ones we are dealing with, it is not unexpected that more observations are needed to arrive at reliable estimates (see also the `GJRM`-package manual (p. 22) in the penalized likelihood framework in the $p < n$ case). Second, boosting distributional regression with a unique step length for all parameter submodels might lead to imbalanced updates of predictors in some scenarios as Zhang et al. (2022) outline. In certain situations, this can become a problem as some submodels might not be appropriately fitted within a limited number of boosting iterations. As a result, the authors propose using adaptive step-lengths for univariate Gaussian location-scale models and it could be interesting to test this procedure in other univariate and multivariate GAMLSS family classes. Third, a further limitation of model-based boosting is its computationally expensive tuning procedure based on cross-validation. Alternatives avoiding time-consuming resampling procedures have been proposed based on so-called random probes (Thomas et al., 2017). However, these come at a cost of higher memory demand. Still it may be interesting to investigate such alternatives in our model class.

Another direction for future research, we aim to explore is the extension to boosted copula regression models with discrete and continuous-discrete outcomes to leverage Bayesian and frequentist counterparts (Klein et al., 2019; Radice et al., 2016). Moreover, it would be interesting to analyze multivariate responses of higher dimension. While such a setting can conceptually be embedded into distributional copula regression in a straightforward manner, model building and selection of all marginal distributions and a parametric copula as well as stable estimation are far more challenging. Here, it seems to be more promising to resort to semiparametric methods not requiring parametric assumptions on the margins and the dependence structure (e.g., Klein et al., 2022, and reference therein) or other forms of copulas such as pair-copula constructions (Vatter & Nagler, 2018).

ACKNOWLEDGMENTS

Nicolai Hans, Nadja Klein, and Andreas Mayr gratefully acknowledge support by the German research foundation (DFG) through the grants KL3037/2-1, MA7304/1-1 (428239776).

DATA AVAILABILITY STATEMENT

Publication or distribution of the original data is not allowed following legal restrictions on protection of personal data.

ORCID

Nadja Klein https://orcid.org/0000-0002-5072-5347

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

Andreas Mayr https://orcid.org/0000-0001-7106-9732

REFERENCES

Barker, D.J. (1997) The long-term outcome of retarded fetal growth. *Clinical Obstetrics and Gynecology*, 40(4), 853–863.

Bermingham, M.L., Pong-Wong, R., Spiliopoulou, A. et al. (2015) Application of high-dimensional feature selection: evaluation for genomic prediction in man. *Scientific Reports*, 5(10312), 1–12.

Boulet, S.L., Alexander, G.R., Salihu, H.M. & Pass, M. (2003) Macrosomic births in the united states: determinants, outcomes, and proposed grades of risk. *American Journal of Obstetrics and Gynecology*, 188(5), 1372–1378.

Bühlmann, P. & Hothorn, T. (2007) Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22(4), 477–505.

Bühlmann, P. & Yu, B. (2003) Boosting with the L_2 loss: regression and classification. *Journal of the American Statistical Association*, 98(462), 324–339.

Craiu, V.R. & Sabeti, A. (2012) In mixed company: Bayesian inference for bivariate conditional copula models with discrete and continuous outcomes. *Journal of Multivariate Analysis*, 110, 106–120.

Dudley, N. (2005) A systematic review of the ultrasound estimation of fetal weight. *Ultrasound in Obstetrics and Gynecology*, 25(1), 80–89.

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

Espasandín-Domínguez, J., Cadarso-Suárez, C., Kneib, T., Marra, G., Klein, N., Radice, R. et al. (2019) Assessing the relationship between markers of glycemic control through flexible copula regression models. *Statistics in Medicine*, 38(27), 5161–5181.

Fahrmeir, L., Kneib, T., Lang, S. & Marx, B. (2013) *Regression—models, methods and applications*. Berlin: Springer.

Faschingbauer, F., Beckmann, M., Goecke, T., Yazdi, B., Siemer, J., Schmid, M. et al. (2012) A new formula for optimized weight estimation in extreme fetal macrosomia (≥ 4500 g). *Ultraschall in der Medizin*, 33(5), 480–488.

Faschingbauer, F., Dammer, U., Raabe, E., Schneider, M., Faschingbauer, C., Schmid, M. et al. (2015) Sonographic weight estimation in fetal macrosomia: influence of the time interval between estimation and delivery. *Archives of Gynecology and Obstetrics*, 292, 59–67.

Faschingbauer, F., Dammer, U., Raabe, E., Kehl, S., Schmid, M., Schild, R.L. et al. (2016) A new sonographic weight estimation formula for small-for-gestational-age fetuses. *Journal of Ultrasound in Medicine*, 35(8), 1713–1724.

Faschingbauer, F., Raabe, E., Heimrich, J., Faschingbauer, C., Schmid, M., Mayr, A. et al. (2016) Accuracy of sonographic fetal weight estimation: influence of the scan-to-delivery interval in combination with the applied weight estimation formula. *Archives of Gynecology and Obstetrics*, 294(3), 487–493.

Freund, Y. & Schapire, R.E. (1996) Experiments with a new boosting algorithm. In: *Proceedings of the thirteenth international conference on machine learning theory*. San Francisco: Morgan Kaufmann, pp. 148–156.

Friedman, J., Hastie, T. & Tibshirani, R. (2000) Additive logistic regression: a statistical view of boosting. *The Annals of Statistics*, 28(2), 337–374.

Friedman, J.H. (2001) Greedy function approximation: a gradient boosting machine. *The Annals of Statistics*, 29(5), 1189–1232.

Ghosh, S. (2014) Multivariate analyses of blood pressure related phenotypes in a longitudinal framework: insights from GWAS 18. *Genetic Epidemiology*, 38(Suppl 1), S63–S67.

Gneiting, T. & Raftery, A.E. (2007) Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477), 359–378.

Hastie, T. (2007) Comment: boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22(4), 513–515.

Hastie, T., Tibshirani, R. & Friedman, J. (2009) *The elements of statistical learning: data mining, inference and prediction*, 2nd edition, New York: Springer.

Hastie, T.J. & Tibshirani, R.J. (1990) *Generalized additive models*. London: CRC.

Hofner, B., Kneib, T. & Hothorn, T. (2016) A unified framework of constrained regression. *Statistics and Computing*, 26(1), 1–14.

Hofner, B., Mayr, A., Robinzonov, N. & Schmid, M. (2014) Model-based boosting in R: a hands-on tutorial using the R package mboost. *Computational Statistics*, 29, 3–35.

Hofner, B., Mayr, A. & 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.

Hothorn, T., Leisch, F., Zeileis, A. & Hornik, K. (2005) The design and analysis of benchmark experiments. *Journal of Computational & Graphical Statistics*, 14(3), 675–699.

Jordan, A., Krüger, F. & Lerch, S. (2019) Evaluating probabilistic forecasts with scoring rules. *Journal of Statistical Software*, 90(12), 1–37.

Klein, N., Hothorn, T., Barbanti, L. & Kneib, T. (2022) Multivariate conditional transformation models. *Scandinavian Journal of Statistics*, 49(1), 116–142.

Klein, N. & Kneib, T. (2016) Simultaneous inference in structured additive conditional copula regression models: a unifying Bayesian approach. *Stat. and Computing*, 26(4), 841–860.

Klein, N., Kneib, T., Marra, G., Radice, R., Rokicki, S. & McGovern, M.E. (2019) Mixed binary-continuous copula regression models with application to adverse birth outcomes. *Statistics in Medicine*, 38(3), 413–436.

Kolev, N. & Paiva, D. (2009) Copula-based regression models: a survey. *Journal of Statistical Planning and Inference*, 139(11), 3847–3856.

Marra, G. & Radice, R. (2017) Bivariate copula additive models for location, scale and shape. *Computational Statistics & Data Analysis*, 112, 99–113.

Mayr, A., Fenske, N., Hofner, B., Kneib, T. & 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*, 61(3), 403–427.

Mayr, A., Hofner, B. & Schmid, M. (2012) The importance of knowing when to stop. *Methods of Information in Medicine*, 51(02), 178–186.

McNeil, A.J., Frey, R. & Embrechts, P. (2005) *Quantitative risk management: concepts, techniques and tools*. New Jersey: Princeton University Press.

Nelsen, R.B. (2006) *An introduction to copulas*, 2nd edition, New York: Springer.

Ott, J. & Wang, J. (2011) Multiple phenotypes in genome-wide genetic mapping studies. *Protein & Cell*, 2(7), 519–522.

Patton, A.J. (2006) Modelling asymmetric exchange rate dependence. *International Economic Review*, 47(2), 527–556.

Petterle, R.R., Laureano, H.A., da Silva, G.P. & Bonat, W.H. (2021) Multivariate generalized linear mixed models for continuous

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024

bounded outcomes: analyzing the body fat percentage data. *Statistical Methods in Medical Research*, 30(12), 2619–2633.

R Core Team (2020) *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. R version 4.0.3.

Radice, R., Marra, G. & Wojtyś, M. (2016) Copula regression spline models for binary outcomes. *Statistics and Computing*, 26(5), 981–995.

Rigby, R.A. & Stasinopoulos, M.D. (2005) Generalized additive models for location, scale and shape. *Journal of the Royal Statistical Society: Series C*, 54(3), 507–554.

Romero, R., Espinoza, J., Gotsch, F., Kusanovic, J.P., Friel, L., Erez, O. et al. (2006) The use of high-dimensional biology (genomics, transcriptomics, proteomics, and metabolomics) to understand the preterm parturition syndrome. *British Journal of Obstetrics and Gynaecology*, 113, 118–135.

Sabeti, A., Wei, M. & Craiu, R.V. (2014) Additive models for conditional copulas. *Statistics*, 3(1), 300–312.

Schild, R., Maringa, M., Siemer, J., Meurer, B., Hart, N., Goecke, T., et al. (2008) Weight estimation by three-dimensional ultrasound imaging in the small fetus. *Ultrasound in Obstetrics and Gynecology*, 32(2), 168–175.

Sklar, M. (1959) Fonctions de répartition à *n* dimensions et leurs marges. *Publications de l'Institut de Statistique de l'Université de Paris*, 8, 229–231.

Song, P.X. (2000) Multivariate dispersion models generated from Gaussian copula. *Scandinavian Journal of Statistics*, 27(2), 305–320.

Staerk, C. & Mayr, A. (2021) Randomized boosting with multivariable base-learners for high-dimensional variable selection and prediction. *BMC Bioinformatics*, 22(1), 1–28.

Strömer, A., Staerk, C., Klein, N., Weinhold, L., Titze, S. & 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.

Thomas, J., Hepp, T., Mayr, A. & Bischl, B. (2017) Probing for sparse and fast variable selection with model-based boosting. *Computational and Mathematical Methods in Medicine*, 2017, 1–8.

Thomas, J., Mayr, A., Bischl, B., Schmid, M., Smith, A. & Hofner, B. (2018) Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Statistics and Computing*, 28, 673–687.

Vatter, T. & Nagler, T. (2018) Generalized additive models for pair-copula constructions. *Journal of Computational and Graphical Statistics*, 27(4), 715–727.

Villar, J., Ismail, L.C., Victora, C.G., Ohuma, E.O., Bertino, E., Altman, D.G. et al. (2014) International standards for newborn weight, length, and head circumference by gestational age and sex: the newborn cross-sectional study of the INTERGROWTH-21st Project. *The Lancet*, 384(9946), 857–868.

Wood, S.N. (2017) *Generalized additive models: an introduction with R*, 2nd edition, London: Chapman & Hall.

Yan, J., Li, T., Wang, H., Huang, H., Wan, J., Nho, K., et al. (2015) Cortical surface biomarkers for predicting cognitive outcomes using group $l_{2,1}$ norm. *Neurobiology of Aging*, 36(Suppl 1), S185–S193.

Zhang, B., Hepp, T., Greven, S. & Bergherr, E. (2022) Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Computational Statistics*, 37, 2295–2332.

SUPPORTING INFORMATION

Web Tables and Figures referenced in Sections 2–5, along with an R implementation as well as code for the simulation studies, are available with this paper at the Biometrics website on Wiley Online Library.

How to cite this article: Hans, N., Klein, N., Faschingbauer, F., Schneider, M. & Mayr, A. (2023) Boosting distributional copula regression. *Biometrics*, 79, 2298–2310. https://doi.org/10.1111/biom.13765

Downloaded from https://academic.oup.com/biometrics/article/79/3/2298/7513798 by G2M Cancer Drugs AG user on 16 October 2024