Received: 6 May 2022 | Revised: 3 February 2023 | Accepted: 17 February 2023

DOI: 10.1002/sim.9699

RESEARCH ARTICLE

Statistics in Medicine WILEY

Boosting multivariate structured additive distributional regression models

Annika Strömer[1] | **Nadja Klein**[2] | **Christian Staerk**[1] | **Hannah Klinkhammer**[1,3] | **Andreas Mayr**[1]

[1]Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, Bonn, Germany

[2]Chair of Uncertainty Quantification and Statistical Learning, Research Center Trustworthy Data Science and Security (UA Ruhr) and Department of Statistics (Technische Universität Dortmund), Dortmund, Germany

[3]Institute for Genomic Statistics and Bioinformatics, University Hospital Bonn, Bonn, Germany

Correspondence
Annika Strömer, Department of Medical Biometrics, Informatics and Epidemiology, University Hospital Bonn, 53127 Bonn, Germany.
Email: stroemer@imbie.uni-bonn.de

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: Grant/Award Numbers: 428239776, KL3037/2-1, MA7304/1-1

We develop a model-based boosting approach for multivariate distributional regression within the framework of generalized additive models for location, scale, and shape. Our approach enables the simultaneous modeling of all distribution parameters of an arbitrary parametric distribution of a multivariate response conditional on explanatory variables, while being applicable to potentially high-dimensional data. Moreover, the boosting algorithm incorporates data-driven variable selection, taking various different types of effects into account. As a special merit of our approach, it allows for modeling the association between multiple continuous or discrete outcomes through the relevant covariates. After a detailed simulation study investigating estimation and prediction performance, we demonstrate the full flexibility of our approach in three diverse biomedical applications. The first is based on high-dimensional genomic cohort data from the UK Biobank, considering a bivariate binary response (chronic ischemic heart disease and high cholesterol). Here, we are able to identify genetic variants that are informative for the association between cholesterol and heart disease. The second application considers the demand for health care in Australia with the number of consultations and the number of prescribed medications as a bivariate count response. The third application analyses two dimensions of childhood undernutrition in Nigeria as a bivariate response and we find that the correlation between the two undernutrition scores is considerably different depending on the child's age and the region the child lives in.

KEYWORDS

generalized additive models for location, scale and shape, model-based boosting, multivariate Gaussian distribution, multivariate logit model, multivariate Poisson distribution, semiparametric regression

1 | INTRODUCTION

Many modern regression models relate certain characteristics of a univariate response distribution to explanatory variables. Examples include generalized additive models (GAMs)[1,2] and quantile regression models,[3] where with the former

the conditional expectation and with the latter conditional quantiles of a univariate response distribution are modeled by an additive decomposition of different covariate effects. In biomedical research there are often multiple endpoints that are typically analyzed separately by univariate regression models where each endpoint serves once as response variable.[4] However, in practice, the components of a multivariate response are often not (conditionally) independent, so that separate models might induce a loss of information and could even lead to potentially misleading conclusions.

A well-known approach for the analysis of multivariate responses, particularly common in the economics literature, is called seemingly unrelated regression.[5] This classical approach is restricted to linear predictors and constant covariance matrices not depending on the covariates; however, extensions to semiparametric predictors for the marginal means exist.[6] Beyond that, multiple discrete responses (eg, count data) can be analyzed using seemingly unrelated Poisson regression[7] and non-linear predictors.[8] Similar to the approach of Zellner[5] these models are limited in their flexibility, and only the expected value of the response is linked to the covariates.[9] A more flexible framework is provided by generalized additive models for location, scale and shape (GAMLSS),[10] in which each parameter of the conditional distribution is modeled by an additive predictor. The use of additive predictors for all distribution parameters, such as location, scale or skewness parameters allows to incorporate different effect types for the covariates in a very flexible way. Klein et al.[11] extended this framework for multivariate responses to model the joint distribution of two or more responses in the spirit of GAMLSS relying on a fully Bayesian approach.

In the medical literature, one common application of GAMLSS for univariate responses is the construction of reference growth charts,[12,13] where also the World Health Organization recommends to use GAMLSS.[14] As the complete conditional distribution is modeled based on covariates (eg, the age of a child), the corresponding quantiles can nicely adapt to a covariate-specific skewness. For multivariate responses, these growth charts could hence not only be constructed separately for different biometrical parameters,[13] but also jointly for multiple characteristics.[15]

In high-dimensional data situations where the number of predictors exceeds the number of observations ($p > n$), classical estimation approaches are no longer directly feasible for our multivariate distributional regression settings. A few exceptions exist where Bayesian variable selection[16] and penalized regression methods[17,18] have been proposed. Nevertheless, in terms of software implementation, GAMLSS based on penalized likelihood estimation is currently only available for univariate response variables.

An alternative fitting approach is statistical boosting, which was originally developed in the field of machine learning and later extended to statistical modeling.[19,20] Its main features are the great flexibility regarding the effect types (eg, spatial, smooth, or random effects) and the data-driven variable selection mechanism. The latter can be particularly useful when the focus is on obtaining sparse models for a possibly high-dimensional covariate space.[21] The concept of boosting has already been extended to distributional regression leading to an algorithm that is able to estimate and select additive predictors for all distribution parameters in univariate GAMLSS.[22,23]

In this work, we adapt the boosting algorithm for multivariate responses by combining the properties of GAMLSS and the main features of statistical boosting. Due to the structure of the algorithm, our approach is able to simultaneously model all distribution parameters and to select possible predictor effects in multivariate distributional regression models: The new multivariate boosting approach allows to model not only the marginals but also the associations between multiple outcomes through additive predictors without requiring the manual selection or comparison of different models. The application of our approach is particularly suitable for exploratory analyses (hypothesis generating) where data-driven variable selection is of special interest.

Motivated by three biomedical applications, we focus on modeling and investigating specific bivariate regression models with emphasis on the most common parametric distributions in biomedical research: the bivariate Bernoulli distribution for binary outcomes, the bivariate Poisson distribution for count data and the bivariate Gaussian distribution for continuous outcomes.[24–26]

In the first biomedical application, the joint genetic predisposition for chronic ischemic heart disease and high cholesterol is analyzed based on a large cohort data from the UK Biobank[27] via the bivariate Bernoulli distribution. The main interest is to study the dependence of these phenotypes on the genetic variants and to discover possible joint associations of the two outcome variables, which is not feasible via classical approaches modeling the phenotypes separately.[28] In our case, we want to gain deeper insights into the relationship between chronic ischemic heart disease and high cholesterol, and the genetic variants affecting their association.

In the second application, we investigate effects for the demand on health care in Australia reported by Cameron and Trivedi.[29] The two considered outcomes are the number of consultations with a doctor and the number of prescribed medications, whose association is modeled using the bivariate Poisson distribution for the covariates gender, age and annual income. The research question is based on a previous analysis by Karlis and Ntzoufras[30] however we illustrate that our approach offers higher flexibility.

In the last epidemiological application, two indicators for acute and chronic undernutrition of children in Nigeria are jointly analyzed, which is motivated by a previous analysis by Klein et al.[11] The two scores are modeled with a bivariate Gaussian distribution, in which besides the marginal expectations also the scale parameter and the correlation parameter depend on covariates. In addition to several covariates describing the life situation of the children, the mother and the household they are living, spatial effects based on regional information are incorporated.

The structure of this article is as follows: Section 2 starts with a brief introduction to multivariate distributional regression models. Then we investigate the different bivariate regression models and give an insight into statistical boosting and the extended algorithm. In Section 3, we illustrate different data settings using a simulation study while Section 4 illustrates the application on biomedical research questions for the considered distributional regression models in Section 2.

2 | BOOSTING MULTIVARIATE DISTRIBUTIONAL REGRESSION

2.1 | The notion of multivariate distributional regression models

In multivariate structured additive distributional regression[11] it is assumed that the conditional distribution $\mathbb{P}_{\boldsymbol{Y}|\boldsymbol{X}=\boldsymbol{x}}$ of a D-dimensional vector of responses $\boldsymbol{Y} = (Y_1, \ldots, Y_D)^\top$ given covariate information summarized in $\boldsymbol{X} = \boldsymbol{x}$ has a K-parametric density $p(\boldsymbol{y}|\boldsymbol{x}) = p(\boldsymbol{y}|\boldsymbol{\theta}(\boldsymbol{x}))$ with covariate dependent distribution parameters $\boldsymbol{\theta}(\boldsymbol{x}) \equiv \boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\top$.

Each distribution parameter θ_k is linked to a structured additive predictor η_k[31] via bijective parameter-specific link functions g_k, such that $g_k(\theta_k) = \eta_k$ and $g_k^{-1}(\eta_k) = \theta_k$, $k = 1, \ldots, K$. The inverse link functions $g_k^{-1} \equiv h_k$ are called response functions and ensure potential restrictions of the parameter space of θ_k. The additive predictors η_k depend on (possibly different) subsets of $\boldsymbol{x}$ and are of the form

$$g_k(\theta_k) = \eta_k = \beta_{0k} + \sum_{j=1}^{p_k} f_{jk}(\boldsymbol{x}), \quad \text{for } k = 1, \ldots, K,$$

where β_{0k} are the intercepts and each $f_{jk}, j = 1, \ldots, p_k$, represents the functional effect of covariates $\boldsymbol{x}$. The effects of the covariates can be specified in a very flexible manner and can correspond to linear, non-linear, random, interaction and further effects.[2,32] Motivated by our applications in Section 4, in this work we focus on the following effect types:

1. Linear effects are represented by $f_{jk}(\boldsymbol{x}) = \boldsymbol{x}_{jk}^T \boldsymbol{\beta}_{jk}$, where $\boldsymbol{\beta}_{jk}$ are the regression coefficients and $\boldsymbol{x}_{jk}$ is a covariate subset of $\boldsymbol{x}$ for parameter θ_k ($\boldsymbol{x}_{jk}$ can be chosen individually for each parameter θ_k).
2. Non-linear effects can be included using smooth functions $f_{jk}(\boldsymbol{x})$. As basis functions we use B-Splines with second order difference penalties.[33]
3. Spatial effects based on observations assigned to discrete regions are incorporated using Markov random fields for modeling neighborhood structures $f_{jk}(\boldsymbol{x}) = f_{jk}(s_i)$, where s_i denotes the region s_i observation i is located in Reference 34.

2.2 | Examples of relevant response distributions

In the following, we describe three common bivariate parametric distributions for binary, count and continuous responses, the bivariate Bernoulli, the bivariate Poisson and the bivariate Gaussian distribution. While there are of course other multivariate distributions for discrete and continuous data,[26,35] these three bivariate distributions are arguably most commonly used and are also relevant for our applications.

2.2.1 | Bivariate Bernoulli distribution

For analyzing potentially correlated binary variables $\boldsymbol{Y} = (Y_1, Y_2)^T$, we consider the bivariate Bernoulli distribution with joint probability mass function

$$p(y_1, y_2) = p_{00}^{(1-y_1)(1-y_2)} p_{10}^{y_1(1-y_2)} p_{01}^{(1-y_1)y_2} p_{11}^{y_1 y_2}, \quad y_1, y_2 \in \{0, 1\},$$

TABLE 1 Contingency table

		Y_2		
		0	1	
Y_1	0	p_{00}	p_{01}	$1-p_1$
	1	p_{10}	p_{11}	p_1
		$1-p_2$	p_2	1

where $p_{ij} = P(Y_1 = i, Y_2 = j)$, $i,j \in \{0,1\}$ are the joint probabilities. Then, the contingency table with marginal probabilities $p_d = \mathrm{P}(Y_d = 1)$, $d = 1,2$ is given in Table 1. In a bivariate logistic regression model (logit model), the marginal probabilities $p_1 = \mathrm{P}(Y_1 = 1)$ and $p_2 = \mathrm{P}(Y_2 = 1)$, as well as the odds ratio $\psi = \frac{p_{00}p_{11}}{p_{01}p_{10}}$, can be estimated considering several covariates.[36,37] If Y_1 and Y_2 are independent, then the odds ratio $\psi = 1$. The different additive predictors in the bivariate logit model are

$$\mathrm{logit}(p_i) = \eta_{p_i}, \quad \text{for } i = 1,2 \quad \text{and} \quad \log(\psi) = \eta_\psi.$$

The joint probability p_{11} can be determined from the marginal probabilities p_1, p_2 and the odds ratio ψ via

$$p_{11} = \begin{cases} \frac{1}{2}(\psi-1)^{-1}\{a - \sqrt{a^2+b}\} & , \ \psi \neq 1 \\ p_1 p_2 & , \ \psi = 1, \end{cases}$$

where $a = 1 + (p_1 + p_2)(\psi - 1)$ and $b = -4\psi(\psi-1)p_1p_2$.[38] The joint probabilities p_{10}, p_{01} and p_{00} can be derived from p_{11} and the marginal probabilities.

An alternative approach for modeling bivariate binary responses is the bivariate probit model. However, in this work we focus on the logit model for two reasons: First, one distribution parameter directly corresponds to the odds ratio, which is easier to interpret and much more common in Biostatistics and biomedical research than the correlation of a latent bivariate response $\boldsymbol{Y}^* \sim N(\boldsymbol{0}, \boldsymbol{\Sigma})$ for a probit model, where $Y_d = 1$ *if* $Y_d^* > 0$ and 0 otherwise, $d = 1,2$ and $\boldsymbol{\Sigma}$ a correlation matrix. Second, the bivariate logit model is computationally favorable since it does not require the latent variables $\boldsymbol{Y}^*$.

2.2.2 | Bivariate Poisson distribution

An important bivariate model for analyzing bivariate count data can be constructed from combining three random variables. If $Z_k, k = 1,2,3$ follow independent Poisson distributions with parameters $\lambda_k > 0$, then the two random variables $Y_1 = Z_1 + Z_3$ and $Y_2 = Z_2 + Z_3$ follow a bivariate Poisson distribution with joint probability function given by

$$p(y_1,y_2) = \exp(-(\lambda_1+\lambda_2+\lambda_3))\frac{\lambda_1^{y_1}}{y_1!}\frac{\lambda_2^{y_2}}{y_2!}\sum_{k=0}^{\min(y_1,y_2)}\binom{y_1}{k}\binom{y_2}{k}k!\left(\frac{\lambda_3}{\lambda_1\lambda_2}\right)^k, \quad y_1,y_2 \in \mathbb{N}_0.$$

The marginals also follow Poisson distributions with expectations $\mathbb{E}(Y_1) = \lambda_1 + \lambda_3$ and $\mathbb{E}(Y_2) = \lambda_2 + \lambda_3$. The parameter λ_3 controls the dependency between Y_1 and Y_2 and corresponds to the covariance $\mathrm{Cov}(Y_1, Y_2) = \lambda_3$. If the variables Y_1 and Y_2 are independent, then $\lambda_3 = 0$ and the bivariate Poisson distribution reduces to the product of two independent Poisson distributions. For further details on the bivariate Poisson distribution, we refer to Kocherlakota and Kocherlakota[25] and Johnson et al.[35]

In a bivariate Poisson regression model, each distribution parameter λ_k, $k = 1,2,3$ can be modeled in terms of several explanatory variables via

$$\log(\lambda_k) = \eta_{\lambda_k}, \qquad k = 1,2,3,$$

where η_k is the corresponding predictor for λ_k.

A drawback of this definition of the bivariate Poisson distribution is its property of modeling only data with positive correlations. An alternative was developed in Lakshminarayana et al.[39] by defining the bivariate Poisson distribution as the product of Poisson marginals with a multiplicative factor. This definition also allows for negative correlations, but results in more difficult interpretations. A further alternative allowing for overdispersion in the marginal distributions is the bivariate negative binomial distribution.[25,30,40]

2.2.3 | Bivariate Gaussian distribution

The bivariate Gaussian distribution is one of the most commonly known distributions for considering two continuous responses. In this case, the random vector is written by $\boldsymbol{Y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the density of $\boldsymbol{Y} = (Y_1, Y_2)^T$ is given by

$$f(y_1, y_2) = \frac{1}{2\pi\sqrt{\det(\boldsymbol{\Sigma})}} \exp\left(-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{\mu})^T\boldsymbol{\Sigma}^{-1}(\boldsymbol{y}-\boldsymbol{\mu})\right), \quad y_1, y_2 \in \mathbb{R},$$

and $\boldsymbol{\mu} = (\mu_1, \mu_2)^T$ and $\boldsymbol{\Sigma} = \text{Cov}(Y_1, Y_2)$ are its mean vector and covariance matrix, respectively. The latter is defined by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$$

with marginal variances $\sigma_1^2 = \text{Var}(Y_1)$ and $\sigma_2^2 = \text{Var}(Y_2)$ and correlation parameter $\rho = \text{Cor}(Y_1, Y_2)$. All parameters of the bivariate Gaussian distribution can be again modeled depending on covariates with parameter specific link-functions:

$$\mu_1 = \eta_{\mu_1}, \quad \mu_2 = \eta_{\mu_2}, \quad \log(\sigma_1) = \eta_{\sigma_1}, \quad \log(\sigma_2) = \eta_{\sigma_2} \quad \text{and} \quad \rho/\sqrt{(1-\rho^2)} = \eta_\rho.$$

For further practical and theoretical details of the bivariate Gaussian distribution, we refer to Kotz et al.[26] When the marginal distributions exhibit heavy tails, the bivariate t-distribution is an attractive alternative to the bivariate normal distribution (see Klein et al.[11] and references therein).

2.3 | Estimation via model-based boosting

Boosting originally arose from the field of supervised machine learning[41] but gained increasing popularity in statistics after the concept was adapted to fit statistical regression models.[19,20] Boosting algorithms are a flexible alternative to classical estimation approaches and have several practical advantages, such as the applicability to high-dimensional data problems and data-driven variable selection.[21,42,43] In the context of regression, there exist different types of boosting algorithms.[21,44] Here, we will focus on a component-wise gradient boosting algorithm with regression-type base-learners, which we refer to as *statistical boosting*.[45,46]

This statistical boosting approach is based on minimizing a pre-specified loss function, which represents the regression problem and typically corresponds to the negative log-likelihood l of the response distribution. In every iteration of the boosting algorithm, so-called base-learners are separately fitted to the negative gradient of the loss function, and the best-performing one is updated to the current estimate. A base-learner in our context is a regression function, and usually corresponds to one specific covariate effect in the additive predictor (eg, a linear model as base-learner leads to a linear effect). An overview of possible base-learners can be found in Hofner et al.[47]

For fitting multivariate distributional regression models, we extend the statistical boosting algorithm for generalized additive models for location, scale and shape[22] to multivariate distributions. A schematic overview of the selection of base-learners in one iteration of the boosting algorithm for multivariate responses can be found in Figure 1.

First, for each additive predictor η_k, $k = 1, \ldots, K$, a set of base-learners $h_1(x_1), \ldots, h_{p_k}(x_{p_k})$ has to be specified in advance. Then, the partial derivative $u = \partial l/\partial\theta_k$ of the negative log-likelihood function l with respect to the different distribution parameters θ_k is calculated and each base-learner is fitted separately to the gradient of the corresponding parameter k. For each parameter, the best performing base-learner j_k^* is determined. After these best-fitting base-learners are selected for each dimension k, only the overall best update (with the highest loss reduction) of all distribution parameters is

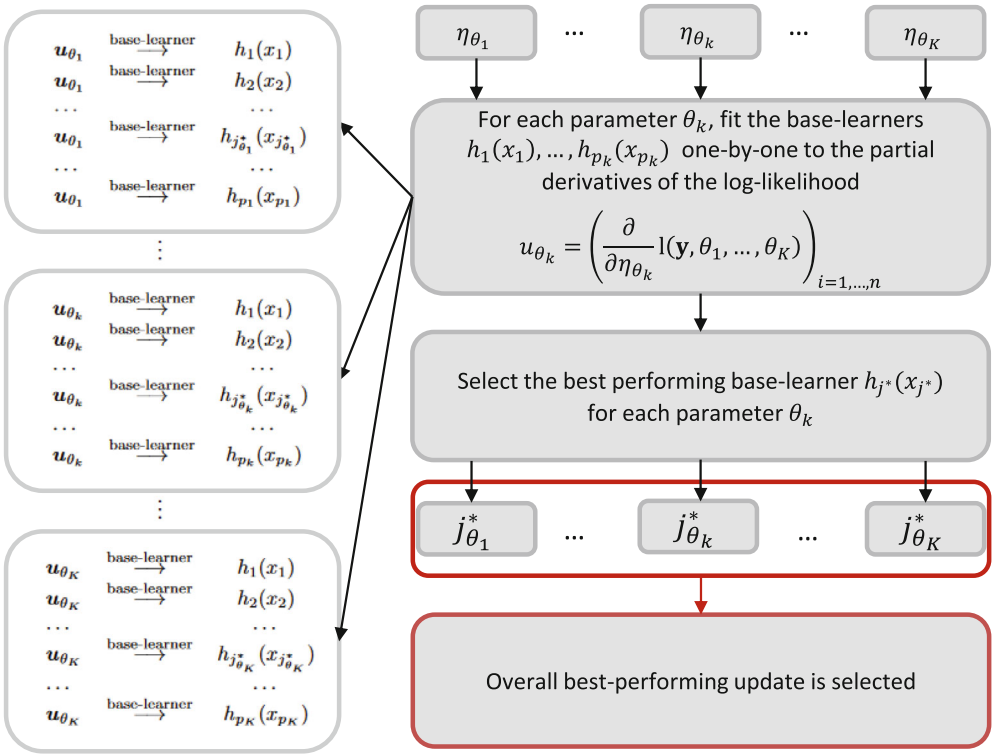

FIGURE 1 Graphical representation of boosting multivariate structured additive distributional regression (displaying one boosting iteration)

finally added to the corresponding additive predictor, with the estimated effect multiplied by a small fixed step-length, for example, $\nu = 0.1$. That means, in every iteration the best-fitting base-learner is determined for each distribution parameter and then compared across the different dimensions. This refers to a so-called non-cyclic version of boosting for distributional regression, leading to a single update of only one distribution parameter in each iteration.[23] Current best-practice in statistical boosting is to use fixed small step-lengths like $\nu = 0.1$ without optimization. Very recently there was work on adaptive step-lengths, particularly for more complex and multi-dimensional models such as GAMLSS. However, so far only the Gaussian location-scale model has been investigated empirically by Zhang et al.[48] The authors propose to use a different step-length for the two parameters, but solutions for more complex models require further research.

The main tuning parameter of the algorithm is the number of boosting iterations, which is typically chosen by cross-validation or resampling techniques. As the algorithm is usually stopped before convergence (*early stopping*), the optimization of the stopping iteration leads to the prevention of overfitting and encourages the sparsity of the resulting model by data-driven variable selection.[49] In particular, those variables, whose corresponding base-learners have never been selected in the update process, are effectively excluded from the final model. The variable selection is simultaneously based on all additive predictors of the corresponding multivariate distribution. The algorithm does not impose any hierarchy between distribution parameters, but only judges the potential predictor variables based on their performance in increasing the joint likelihood. In addition, early stopping typically leads to an improvement in the prediction accuracy and shrinkage of the effect estimates. We provide an implementation of statistical boosting for multivariate distributional regression, which is integrated in the R package **gamboostLSS**.[50]

The boosting approach yields several advantages compared to existing Bayesian and likelihood-based approaches in the context of GAMLSS.[10,11] First, boosting incorporates data-driven variable selection. The issue of variable selection is particularly important in complex model classes, for example, for multivariate distributional regression. The complexity can further increase in settings with many distribution parameters K or high-dimensional predictors with many covariates. In these cases the boosting approach could be favorable since it avoids manual selection of a large number of potential candidate models. Second, the effect estimates are shrunken towards zero due to early stopping of the boosting algorithm. This tends to result in more stable predictions as the variance of the estimates is reduced. Finally, the boosting algorithm can be also applied for high-dimensional data problems, where we have more covariates than observations ($p > n$). Other

approaches, such as more classical Bayesian approaches, are no longer applicable or computationally very demanding for these data situations.

3 | SIMULATIONS

To evaluate the performance of the proposed statistical boosting approach, we conducted a detailed simulation study for the three response distributions presented in Section 2.2. For each distribution, the particular settings are guided by the different applications in Section 4. With our simulations, we aim to answer the following questions:

- Does the boosting approach yield accurate estimates for the corresponding distribution parameters of the bivariate distributions?
- Can the boosting approach identify the truly informative variables and their effects?
- How do the bivariate models perform compared to univariate models that assume independence between the two response components?

In particular, we evaluate the estimation, variable selection and predictive performance. Note that for each considered simulation setting, different variables are informative for the distribution parameters and some of them partially overlap. Therefore, we refer to informative and non-informative variables and do not mention all of them individually for the different settings.

For all simulations, the step-length (learning rate) of the boosting algorithm is set to a fixed value of $\nu = 0.1$ for each parameter of the bivariate models, as well as for the univariate boosted models. The stopping iteration m_{stop} is optimized by minimizing the empirical risk on an additional validation data set with $n_{\text{val}} = 1500$ observations, following the same distribution as the training data. In addition, test data with 1000 observations were generated for the evaluation of the predictive performance (from the same distribution as the training data). As evaluation criteria, multivariate proper scoring rules, namely the negative log-likelihood and the energy score, were used. The energy score generalizes the continuous ranked probability score for multivariate quantities.[51] In addition, univariate distribution-specific evaluation criteria were used, that is, the mean squared error of prediction (MSEP), the area under the curve (AUC) and the Brier score. The Brier score can be used to assess the accuracy of binary classifications and prediction models and is similar to the mean squared error of prediction by considering the mean squared difference between the actual binary outcome and its predicted probability.[52] In contrast to the AUC, which basically only measures the discriminatory power, the Brier score additionally also considers the calibration of the prediction model. Note that the AUC and Brier score do not account for the dependence between the two outcomes and are calculated separately for both outcomes, while the negative log-likelihood and energy scores are probabilistic measures considering the entire joint outcome distribution. A total of 100 simulation runs were performed for each simulation setting.

The corresponding R code to reproduce the results is available on GitHub https://github.com/AnnikaStr/DistRegBoost. Further simulation results, such as comparisons with seemingly unrelated regression and Bayesian approaches, can be found in the Appendix.

3.1 | Bivariate Bernoulli distribution

3.1.1 | Simulation design

For the simulation of the bivariate logit model, we considered a situation with $n = 1000$ observations and $p = 1000$ covariates for each of the three distribution parameters, which corresponds to a high-dimensional situation as the number of possible regression coefficients has tripled due to the three distribution parameters ($3p > n$). For data generation, the R package **VGAM**[53] was used, whereby the parameters p_1, p_2 and ψ were simulated with the following linear predictors

$$\text{logit}(p_1) = \eta_{\mu_1} = X_1 + 1.5X_2 - X_3 + 1.5X_4, \qquad \text{logit}(p_2) = \eta_{\mu_2} = 2X_1 - X_2 + 1.5X_3,$$
$$\log(\psi) = \eta_{\psi} = -1.5 + 1X_3 + 1.5X_4.$$

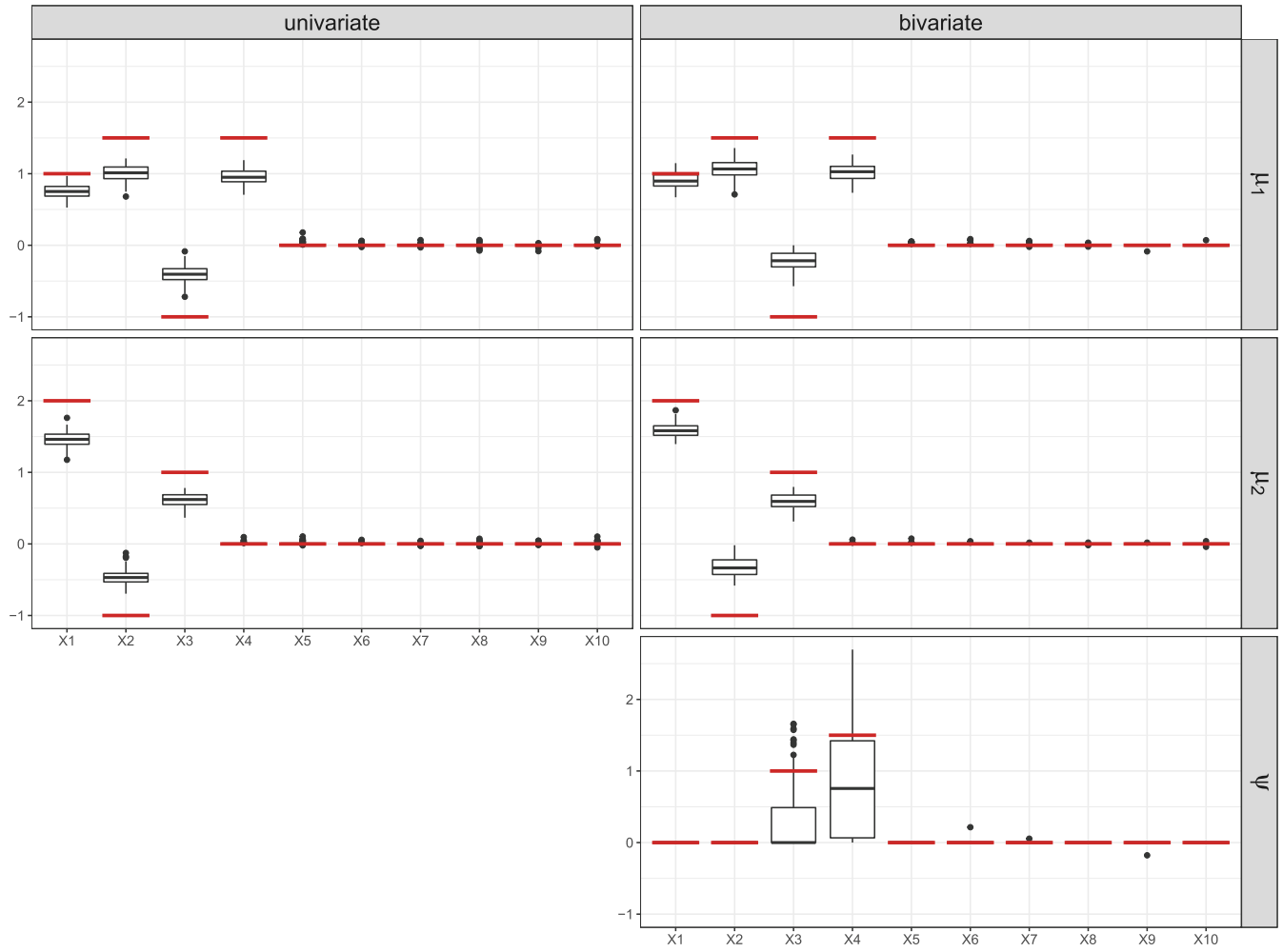

FIGURE 2 Results for the estimated linear effects of the univariate (left) and bivariate Bernoulli (right) model of the first ten covariates $X_1, \ldots, X_{10}$ from 100 simulation runs. The red horizontal lines correspond to the true values

Overall, only the first six covariates out of the $p = 1000$ had a relevant effect on any of the distribution parameters (four for p_1, three for p_2 and two for ψ). The covariates were simulated from a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with a Toeplitz covariance structure $\Sigma_{ij} = \rho^{|i-j|}$ for $1 \leq i, j \leq p$, where $\rho = 0.5$ is the correlation between consecutive variables X_j and X_{j+1}. The covariates were incorporated in the boosting approach by using simple linear models as base-learners. As measures for the predictive performance, AUC, the Brier score, the negative log-likelihood and energy score were considered.

3.1.2 | Results

Figure 2 presents the coefficient estimates of the first ten covariates $X_1, \ldots, X_{10}$ in form of boxplots for the univariate (left) and bivariate (right) model with the red horizontal lines corresponding to the true values. The univariate and bivariate models reflect the true structure for η_{μ_1} and η_{μ_2}, as well as η_ψ for the bivariate model, with both models leading to very similar results. The informative variables for μ_1 and μ_2 were correctly selected in almost every simulation run. Specifically, we obtained overall selection rates (averaged over all informative variables) of 100% for the univariate models and of 100% for μ_1 and 98.75% for μ_2 in the bivariate model. The selection rate for ψ is slightly lower than for the other parameters with a selection rate of 59.5% (see Appendix Table A3). The non-informative variables were selected very rarely overall, resulting in sparse models and accurate model specifications that are able to recover the ground truth.

TABLE 2 Resulting predictive performance on independent test data for the linear setting of the bivariate Bernoulli distribution; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Univariate	Bivariate
AUC (Y_1)	0.88 (0.01)	0.88 (0.01)
AUC (Y_2)	0.84 (0.01)	0.84 (0.01)
Brier score (Y_1)	0.14 (0.01)	0.14 (0.01)
Brier score (Y_2)	0.16 (0.01)	0.16 (0.01)
Energy score	0.28 (0.21)	0.27 (0.01)
Negative log-likelihood	930.51 (24.24)	906.64 (29.24)

A comparison of the predictive performance is provided in Table 2, showing that the univariate and bivariate models were very similar in terms of AUC, Brier score, and energy score, with the bivariate model having slightly better negative log-likelihood. In addition, the energy score for the univariate models showed a larger standard deviation. Further simulation results of this linear setting for a low-dimensional data situation ($p = 10$ and $n = 1000$) can be found in Appendix A.1.

3.2 | Bivariate Poisson distribution

3.2.1 | Simulation design

For the bivariate Poisson regression model, we investigated both linear and non-linear settings with $p = 10$ covariates and $n = 1000$ observations for each distribution parameter. For the linear setting, the underlying true predictors were specified as

$$\begin{aligned} \log(\lambda_1) = \eta_{\lambda_1} &= -X_1 + 0.5X_2 + 1.5X_3, \qquad \log(\lambda_2) = \eta_{\lambda_2} = 2X_1 - X_3 + 1.5X_4 + X_5, \\ \log(\lambda_3) = \eta_{\lambda_3} &= 0.5X_5 + X_6 - 0.5X_7, \end{aligned} \tag{1}$$

where the covariates followed a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$ with Toeplitz covariance structure and correlation coefficient $\rho = 0.5$. Thus, the first seven covariates were informative for any of the distribution parameter (three for λ_1 and λ_3, four for λ_2). For this setting, simple linear models were incorporated as base-learners. For the non-linear setting, the true additive predictors were given by

$$\begin{aligned} \log(\lambda_1) = \eta_{\lambda_1} &= \sqrt{X_1}X_1, \qquad \log(\lambda_2) = \eta_{\lambda_2} = \cos(2X_2), \\ \log(\lambda_3) = \eta_{\lambda_3} &= \sin X_3, \end{aligned} \tag{2}$$

where the covariates were independently simulated from the uniform distribution $U(0, 1)$ and only one covariate was informative for each of the distribution parameters. As base-learners, we chose P-splines (20 equidistant knots with a second-order difference penalty and four degrees of freedom). The R **extraDistr**[54] package was used to simulate data from the bivariate Poisson regression model.

3.2.2 | Results

Figure 3 displays the coefficient estimates for the linear Poisson regression models (1). The boxplots present the estimated coefficients for the univariate (left) and bivariate models (right). Overall, boosting the bivariate regression model was able to identify the informative variables and to accurately estimate the true effects represented by the red horizontal lines. In comparison, the univariate models for λ_1 and λ_2 resulted in much smaller estimated coefficients. For both models, the informative variables were selected in almost every simulation run: considering λ_1 and λ_2, the

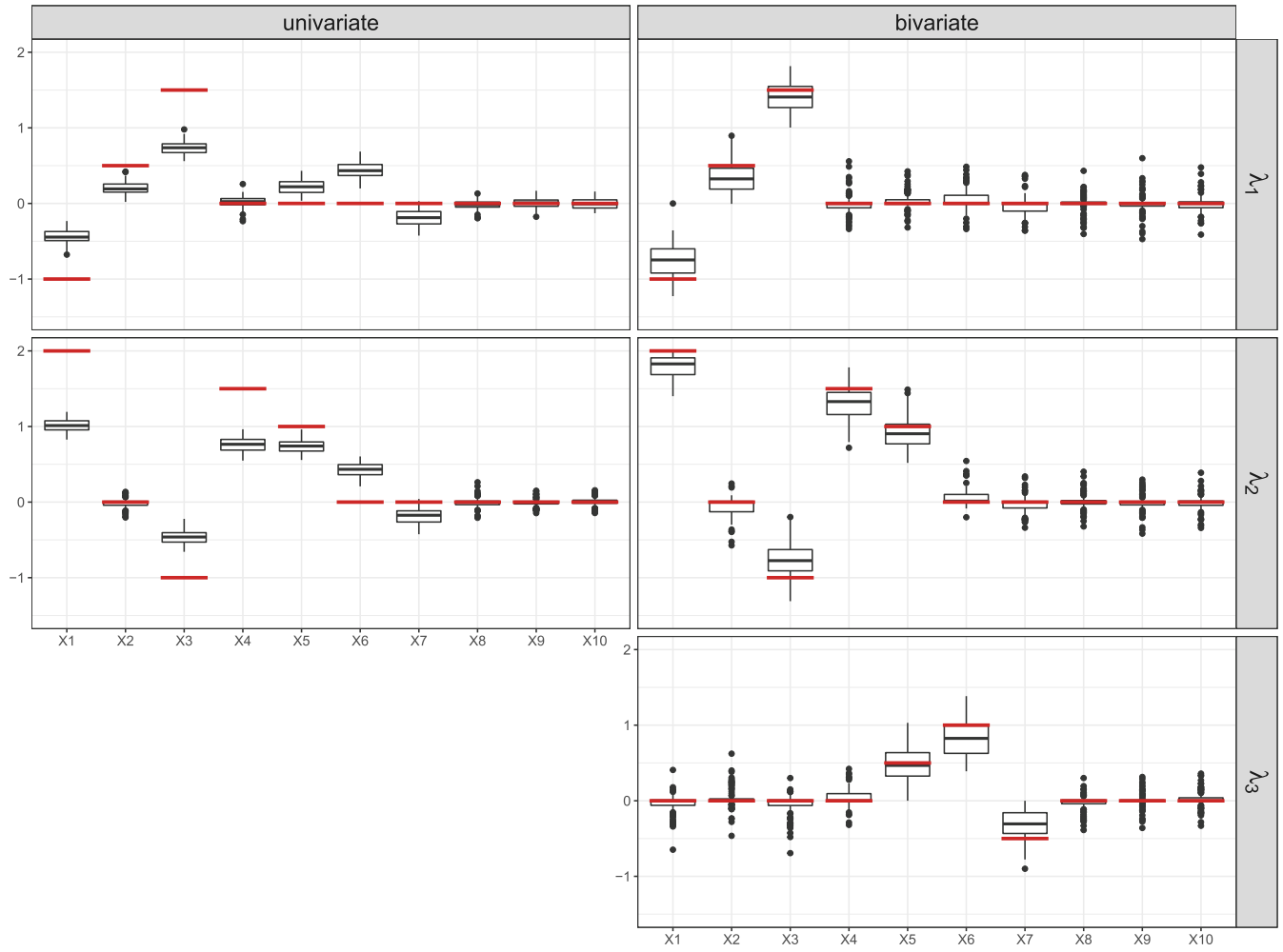

FIGURE 3 Results for the estimated linear effects of the univariate (left) and bivariate Poisson model (right) from 100 simulation runs. The horizontal lines correspond to the true values

univariate models and the bivariate model had a selection rate of almost 100% for the informative variables, whereby also for λ_3 a high selection rate of 95.67% for the informative variables was achieved. On the other hand, the univariate models as well as the bivariate model selected also several non-informative variables with a small coefficient size. A more detailed overview on the selection rates for the specific parameters can be found in Table A8 of the Appendix.

Furthermore, we considered the MSEP, the negative log-likelihood, and the energy score for the evaluation of the predictive performance on test data (see Tables 3 and 4). The MSEP only accounts for the marginal distributions and displays here a slightly better performance for the univariate models. The negative log-likelihood and the energy score, which also take the association into account, showed a better performance for the bivariate model.

Figure 4 displays the effect estimates of the informative variables X_1, X_2 and X_3 for the non-linear setting (2). Overall, the estimated splines approximate the true effects well for each parameter of the bivariate model and clearly outperform the univariate models for λ_1 and λ_2. The informative variables were selected in each simulation run. However, as in the linear model, we observed also high selection rates for the non-informative variables in both models (see Appendix A.2, Table A8).

In terms of predictive performance, similar to the linear setting, the MSEP indicated a better performance of the univariate models, while the bivariate models, as expected, yielded better results in terms of the negative log-likelihood. The energy score was very similar for both models but overall slightly better for the bivariate model. Further simulation results for these settings in case of high-dimensional data with $p = 1000$ covariates and $n = 1000$ observations can be found in Appendix A.2.

TABLE 3 Resulting predictive performance on independent test data for the linear and non-linear settings of the bivariate Poisson regression; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Linear model		Non-linear model	
	Univariate	**Bivariate**	**Univariate**	**Bivariate**
MSEP (Y_1)	2.66 (0.18)	3.96 (0.29)	4.64 (0.25)	8.18 (0.65)
MSEP (Y_2)	2.86 (0.23)	4.11 (0.34)	5.49 (0.29)	9.06 (0.68)
Energy score	1.48 (1.11)	1.36 (0.03)	1.95 (0.04)	1.95 (0.04)
Negative log-likelihood	3598.42 (54.31)	3413.68 (40.91)	4433.06 (52.06)	4246.96 (42.58)

TABLE 4 Resulting predictive performance on independent test data of the bivariate Gaussian regression; mean (SD) values from 100 simulation runs are reported for the univariate and bivariate models

	Univariate	**Bivariate**
MSEP (Y_1)	1.59 (0.11)	1.59 (0.11)
MSEP (Y_2)	1.38 (0.07)	1.38 (0.07)
Energy score	1.03 (0.02)	1.01 (0.02)
Negative log-likelihood	3370.41 (89.59)	3098.11 (109.97)

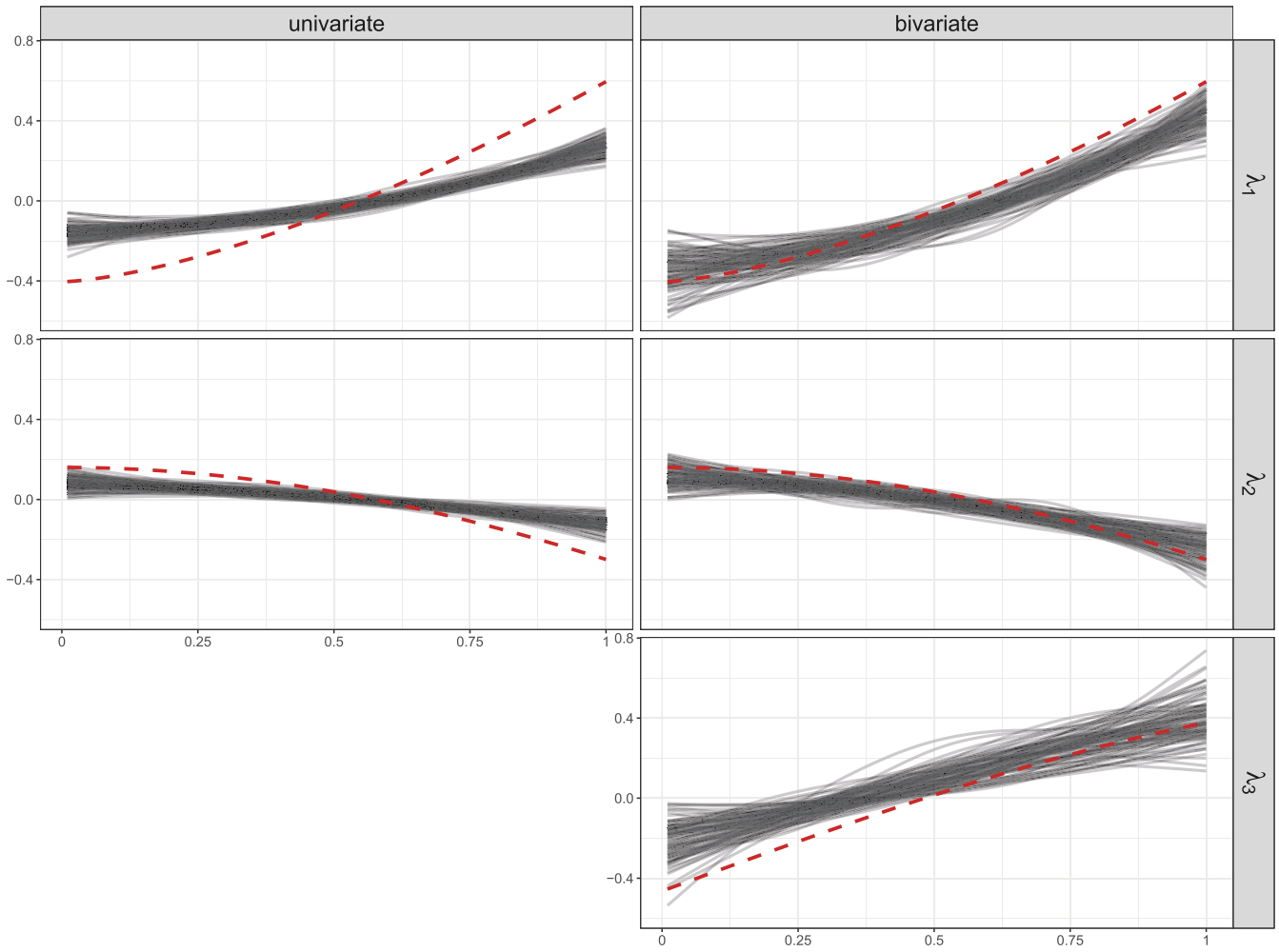

FIGURE 4 Results for the estimated non-linear effects for the univariate (left) and the bivariate Poisson model (right) of the informative variables from 100 simulation runs. The red dotted lines correspond to the true effects

3.3 | Bivariate Gaussian distribution

3.3.1 | Simulation design

For the simulation of a bivariate Gaussian distributed outcome, we considered a setting with linear, non-linear and spatial effects with $p = 10$ covariates and $n = 1000$ observations with the following true predictors

$$\mu_1 = \eta_{\mu_1} = \sin(2X_1)/0.5 + X_6 + 0.5X_7 + f_{\text{spat}} \qquad \mu_2 = \eta_{\mu_2} = 2 + 3\cos(2X_2) + 0.5X_7 + X_8 + f_{\text{spat}}$$
$$\log(\sigma_1) = \eta_{\sigma_1} = \sqrt{X_3}X_3 - 0.5X_8 + f_{\text{spat}} \qquad \log(\sigma_2) = \eta_{\sigma_2} = \cos(X_4)X_4 + 0.25X_9 + f_{\text{spat}}$$
$$\rho/\sqrt{1-\rho^2} = \eta_\rho = \log(X_5^2) + X_{10} + f_{\text{spat}},$$

where the covariates were independently simulated from the uniform distribution $U(0, 1)$. Each included covariate was informative for one of the distribution parameters; more precisely, for each parameter three covariates, one linear and one non-linear, and additionally the spatial effect. For the linear effects $(X_6, \ldots, X_{10})$ we used simple linear models as base-learners and P-splines for the non-linear effects $(X_1, \ldots, X_5)$. The spatial effects were simulated with $f_{\text{spat}}(s) = \sin(x_s^c)\cos(0.5y_s^c), s \in 1, \ldots, S$, based on the centroids of the standardized coordinates of the discrete regions in Western Germany with overall $S = 327$ regions. The neighborhood structure was modeled by the spatial base-learner using a Markov random field based on the `R` package **BayesX**.[55]

3.3.2 | Results

Considering the linear effects $(X_6, \ldots, X_{10})$, the effect estimates for both models reflect the true structures of the linear part of the predictors, whereby the bivariate model better approximates the true values (see Figure A6 in Appendix A.3). The bivariate model was also able to capture the true non-linear functions well (Figure 5); only small deviations are observed for the variance and for the correlation ρ at the left border. The results for the univariate models appear to be very similar regarding the univariate effects and can be found in the Appendix (Figure A7). For the spatial effects, the true structure for the regions in West Germany was identified by each distribution parameter (a graphical representation of the true structure and the estimated spatial effects are in Appendix A.3).

The informative variables for the univariate and bivariate models were selected in nearly all 100 simulation runs, where the bivariate model also correctly selected the informative variables for the correlation between the outcomes. Whereas, we can not examine the correlation with the univariate models. The selection rates for the non-informative variables were slightly higher for the bivariate model (see Appendix Table A11).

Regarding predictive performance, the MSEP, the energy score, and the negative log-likelihood were considered. For the MSEP and the energy score, similar results were observed for the univariate and the bivariate models. The negative log-likelihood on the test set showed an improvement in predictive performance considering the bivariate model. Further simulation results for this setting in case of high-dimensional data with $p = 1000$ covariates and $n = 1000$ observations can be found in Appendix A.3.

3.4 | Summary

Overall, we obtained promising results for all three considered distributional regression families (logistic, Poisson, and Gaussian regression), highlighting that the boosting algorithm yields appropriate estimates for the different parameters and is capable of identifying the most informative variables from a potentially much larger set of candidate variables. The comparison with the univariate models showed that the estimated effects for the bivariate model were able to provide better approximations to the true structure of the predictors than the univariate models (particularly for the Poisson and Gaussian regression models). We noticed that for the logistic regression model, the selection rates for the association parameter, the odds ratio, tended to be lower than for the association parameter of the bivariate Poisson and Gaussian distribution. Furthermore, the number of selected non-informative variables was higher for the univariate as well as the bivariate models for the Poisson distribution. However, the linear low-dimensional setting for the bivariate logistic regression model (see Appendix A.1) and the low-dimensional Gaussian regression model showed higher selection rates in this

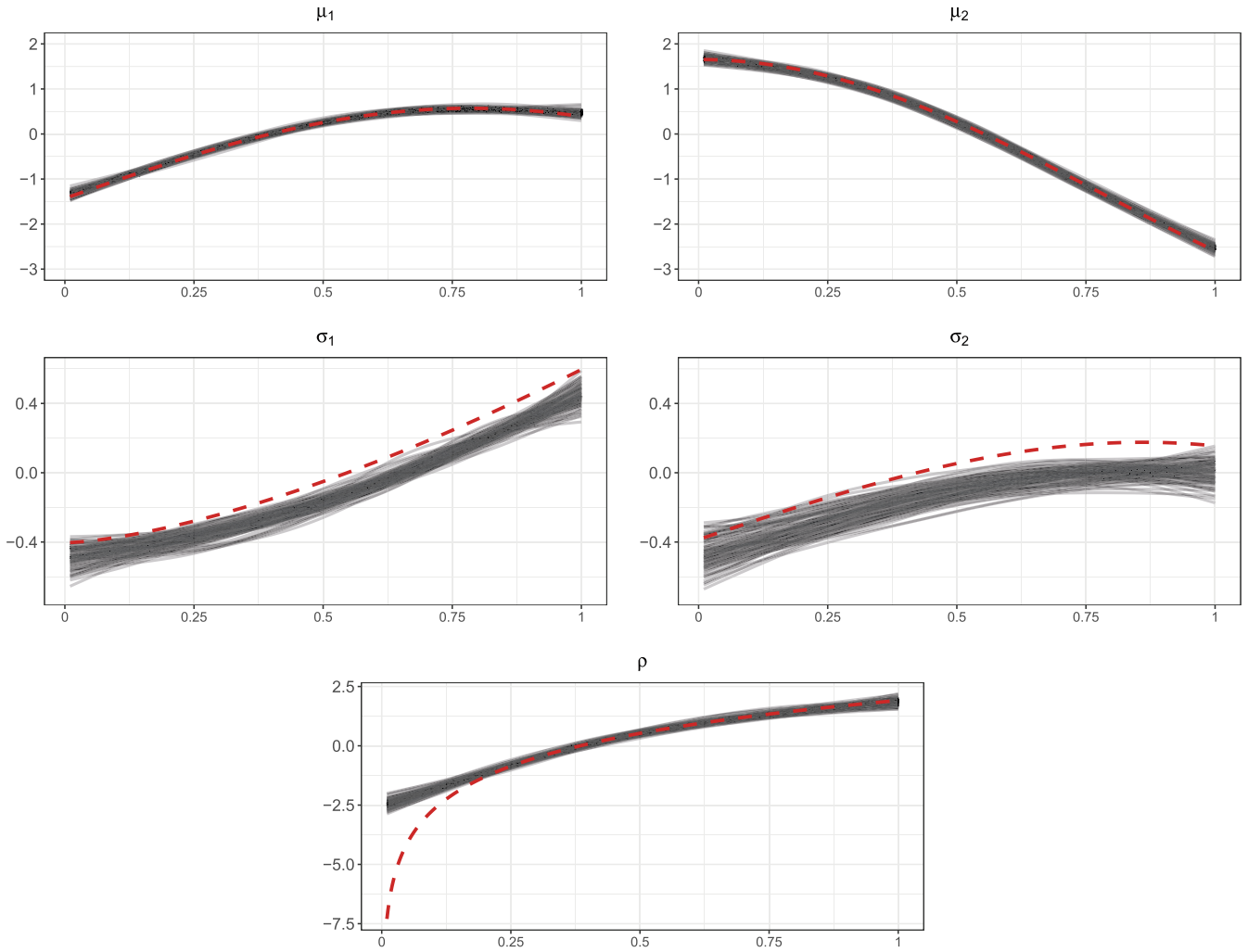

FIGURE 5 Results for the estimated non-linear effects ($X_1, \ldots, X_5$) of the bivariate Gaussian regression model from 100 simulation runs. The red dotted lines correspond to the true effects

situation as well (Appendix A.3). Conclusively, this highlights a tendency of the algorithm to select more non-informative variables in low-dimensional settings.

Regarding prediction accuracy, as expected, the univariate and bivariate models performed similarly for evaluation criteria that consider only the marginals (AUC, Brier score and MSEP). Only for the Poisson distribution, the univariate model performed slightly better regarding the MSEP. This can be explained by the particular design of this bivariate distribution, that is, the summation of the means for both outcomes ($\mathbb{E}(Y_1) = \lambda_1 + \lambda_3$ and $\mathbb{E}(Y_2) = \lambda_2 + \lambda_3$). In Figure 3, for example, we observe that the informative variables X_5, X_6 and X_7 for parameter λ_3 were selected quite frequently with a higher estimated coefficient in the univariate models. These wrongly selected variables for the marginals resulted in an improvement of the MSEP. In the bivariate model, we account for the association between Y_1 and Y_2 by modeling the dependency in terms of the covariates. The MSEP does not account for the association and the variables describing dependency are not reflected in the marginals as in the univariate models. Regarding the predictive scores which account for associations between the outcomes, the energy score tended to be very similar for the univariate and bivariate models, while the negative log-likelihood was consistently better for the bivariate models.

Overall, the bivariate models are obviously more complex because they also model the association between the two outcomes. Regarding the marginals, however, they yield similar models as when analyzing both responses separately. Therefore, we also get similar prediction performance for the marginals from the univariate and bivariate models. The distributional evaluation measures like the negative log-likelihood, on the other hand, are consistently better for the bivariate models as they consider the complete multivariate distribution.

4 | BIOMEDICAL APPLICATIONS

In this section we consider three diverse biomedical data sets to illustrate the applicability of our extended boosting approach for multivariate distributional regression models based on binary, count and continuous outcomes presented in Section 2.2.

4.1 | Genetic predisposition for chronic ischemic heart disease and high cholesterol

For analyzing the association between high cholesterol and chronic ischemic heart disease in dependency of different genetic variants, we used cohort data from the UK Biobank (under application number 81202). The UK Biobank is a large biomedical cohort study containing genetic and health information from over half a million British participants.[27]

In classical approaches for analyzing a potential genetic liability to a specific phenotype such as high cholesterol or chronic heart disease, each considered genetic variant is fitted individually to the phenotype using a simple linear model.[28] In this context, previous works including genome-wide association studies[56,57] have investigated to find genetic variants associated with high cholesterol and heart disease. Using our boosting algorithm for multivariate distributional regression, the main interest here is to investigate the association between chronic ischemic heart disease and high cholesterol, both considered as binary phenotypes (high cholesterol > 6.16 mmol/l). In particular, we aim to identify genetic variants affecting their association by estimating the two phenotypes jointly in a bivariate logistic model. That means we do not only want to model the individual distributions of the two phenotypes, but also estimate the dependency between these phenotypes as a function of genetic variants, which is not possible with conventional approaches.

The considered data set consists of 20,000 randomly sampled observations of individuals with white British ancestry, with additional 10,000 observations used to validate the optimal stopping iteration. The fixed step-length was set to $\nu = 0.1$. For each phenotype, 1000 variants were selected in a pre-screening step based on the largest marginal associations between the variants and the phenotype, which were computed with the PLINK2 function `-variant-score`.[58,59] After pre-screening, the data set contains a total of 1865 variants (with 135 variants selected for both phenotypes). Variants with minor allele frequency not less than 1% were randomly sampled with the `-thin-count` function. Missing genotypes were imputed by the reference allele using the `R` package **bigsnpr**.[60] Note that the pre-screening of 1000 genetic variants for each phenotype and the usage of 20,000 randomly sampled observations from the much larger cohort was performed to avoid computational memory problems. While there exists recent approaches to use boosting to fit multivariable regression models for single phenotypes on the complete data set, classical methods to model the genetic liability are based on summing up univariate effects.[61,62]

Figure 6 shows the resulting estimated coefficients (expressed in exponential absolute values of the estimated coefficients) for the three distribution parameters. When comparing these Manhattan plots with the classical univariate ones (based on the marginal association evaluated on the $-\log 10(P)$ scale) for high cholesterol and chronic ischemic heart disease, we find that the bivariate boosting model tended to identify variants with a higher coefficient value (stronger effect) from similar or the same genomic locations, where the univariate models also showed large univariate associations (see Appendix B.1).

For high cholesterol, for example, the variants with the smallest univariate p-values are located on chromosomes 18 and 19; on these chromosomes there were also the variants that had the highest estimated coefficients in the bivariate boosting model. These findings are consistent with the location of known cholesterol-associated genes.[57] Variants from these chromosomes were also selected with our approach for the odds ratio. Our model selected several variants for chronic ischemic heart disease that are in line with the findings of the meta-analysis of genome-wide association studies examining DNA sequence variants associated with ischemic heart disease of Elosua and Sayols-Baixeras[63] (eg, the variants rs11206510, rs2891168, and rs4420638).

Overall, several variants were selected by the boosting approach for each distribution parameter, that is, 75 for μ_1, 154 for μ_2, and 19 variants for ψ. For the marginal means, mainly those variants were selected that were primarily filtered due to the specific phenotype (the 1000 most highly associated from the univariate screening for both phenotypes). In particular, for μ_1, 63 of the 75 selected variants had been chosen in the pre-screening for ischemic heart disease, so that 12 of the 75 selected variants for μ_1 were primarily selected for high cholesterol. Regarding high cholesterol, 110 variants were selected from the ones that had been pre-selected for this phenotype, while 44 of the selected variants for high cholesterol had been originally pre-selected for ischemic heart disease. The two marginal means μ_1 and μ_2 had six selected variants in common, and both had one variant that was also selected for the odds ratio ψ (namely for μ_1: rs10455872 and for μ_2:

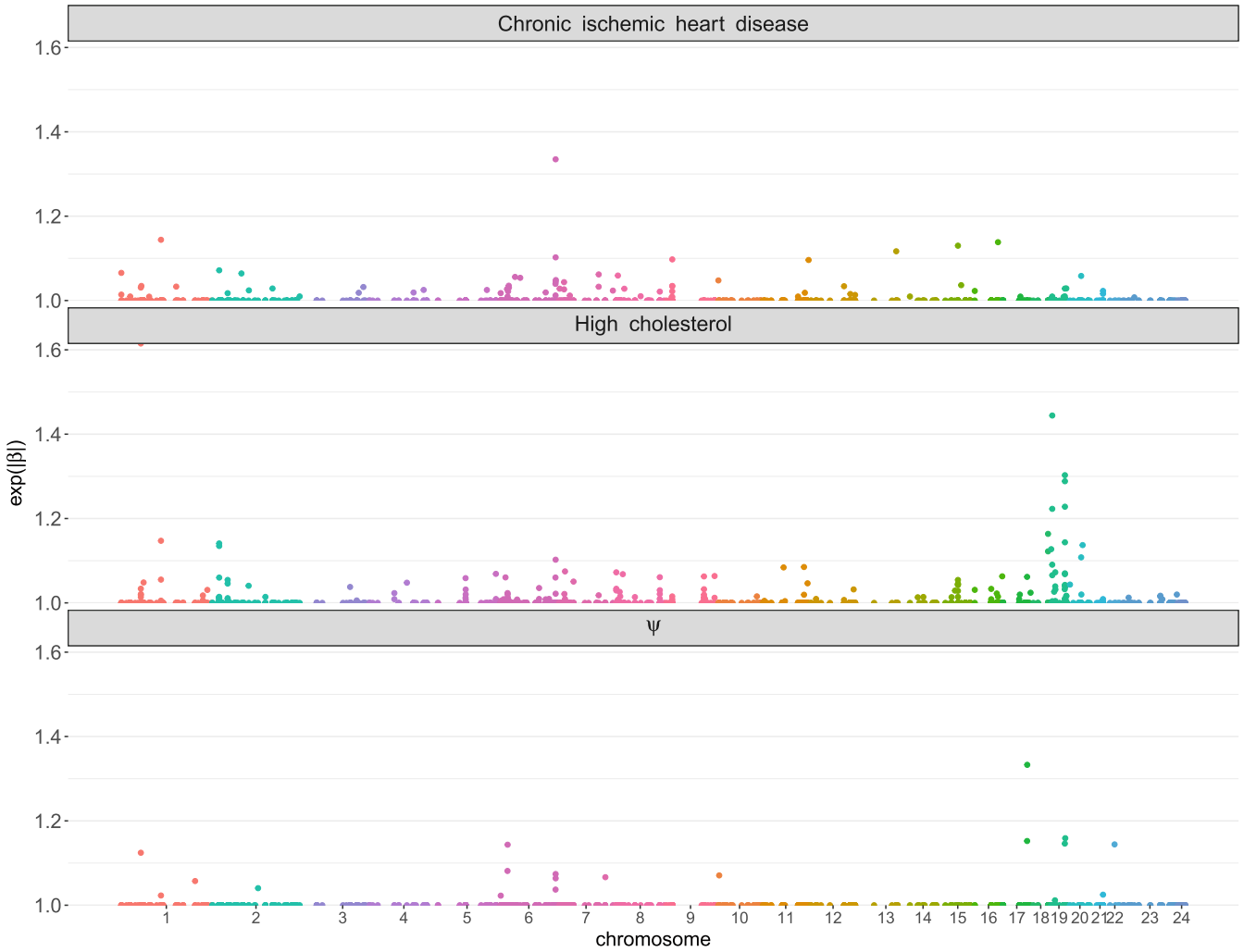

FIGURE 6 Manhattan plots for the coefficient estimates (expressed in exponential absolute values of the estimated coefficients) of the boosted bivariate logistic regression model of the joint analysis of high cholesterol and ischemic heart disease from the UK Biobank data. The *x*-axis represents the genomic location of the variants

rs77542162). The odds ratio included two variants that were among the 1000 most highly correlated pre-selected variants for both phenotypes, namely rs505151 and rs2229094. The other 17 variants selected for the odds ratio were divided as follows: 10 from μ_1 (ischemic heart disease) and 7 from μ_2 (high cholesterol). This means the algorithm identified several variants that affect the dependency between the two phenotypes. The odds ratio is the most common measure for examining the dependency between two binary outcomes in biomedical research and the interpretation in our context is very similar. Thus, the selected variants for the association parameter have an effect on both outcomes, with a positive effect increasing the association between heart disease and high cholesterol and conversely.

In summary, our algorithm provides the ability to study the joint genetic predisposition for chronic ischemic heart disease and high cholesterol. With our approach we can also model the dependence of the association between these two phenotypes on genetic variants, which is not possible with classical approaches. In addition, in line with the literature on cardiovascular genetics, our model selected several variants in genomic regions which had been previously identified to be relevant for the considered phenotypes.

4.2 | Demand for health care in Australia

The first analysis on the demand for health care in Australia, based on the Australian health survey from 1977 to 1978, was reported by Cameron and Trivedi.[29] The considered data set consists of $n = 5190$ observations (which is only a subset

TABLE 5 Results of the bivariate poisson model for the demand of health care for model A and model B (see Figure 7 for the non-linear effect estimates for age and income in model B)

	Covariate	$\lambda_{\text{consulations}}$	$\lambda_{\text{medications}}$	λ_3
Model A	Intercept	−2.10	−2.20	−2.62
	Gender (female)	0.05	0.59	0.61
	Age	1.40	3.29	—
	Income	−0.31	−0.10	—
Model B	Intercept	−2.29	−2.22	−0.35
	Gender (female)	0.13	0.60	0.19

of the overall collected survey). The bivariate count variables of interest are the number of consultations with a doctor (in the past 2 weeks) and the number of prescribed medications (used in the last 2 days), which we model using bivariate Poisson regression. The explanatory variables are *gender* (female coded as 1, male as 0), *age* (in years divided by 100) and annual *income* (in Australian dollars; AUD; divided by 1000, measured as midpoints of coded ranges). More details on the survey and its original analysis can be found in Cameron and Trivedi.[29] The data are provided in the R package **bivpois**,[30] which is available on GitHub (https://github.com/cran/bivpois).

In the following, we use the same representation of the bivariate Poisson distribution as introduced in Section 2.2.2. Each distribution parameter $\lambda_k, k = 1, 2, 3$ is modeled based on explanatory variables. We consider the two following models:

ModelA *Gender*, *age* and *income* are included as covariates for $\lambda_{\text{consulations}}$ (number of doctor consultations) and $\lambda_{\text{medications}}$ (number of medications prescribed), but only *gender* is considered as a covariate for the covariance parameter λ_3 (corresponding to Model (b) in Karlis and Ntzoufras[30]).

ModelB For each model parameter, P-splines are used as base-learners for the continuous variables *age* and *income*, while for *gender* linear effects are used.

The stopping iteration of both models was tuned via 25-fold bootstrapping and the step-length was set to a fixed $\nu = 0.1$.

Considering the results of Model A presented in the upper part of Table 5, we observe that with increasing *age*, both the numbers of doctor consultations and prescribed medications are estimated to increase. *Income* has negative marginal effects on both responses, which means that higher *income* is associated with fewer prescribed medications and fewer doctor appointments. For the covariance parameter λ_3, only *gender* was included as an explanatory variable in Model A. The joint effect of *gender* on the number of doctor consultations and prescribed medications indicates that males and females have different covariance terms. The estimated effect of 0.61 for *gender* suggests that the association between numbers of consultations and medications is higher for women than for men.

The lower part of Table 5 and Figure 7 present the results for Model B. With an increasing *age* up to 50 years, the number of doctor consultations is estimated to increase linearly, with a slight decrease starting around the age of 57 years. The estimated effect of *age* on the number of prescribed medications and the covariance parameter is linear and is negative throughout the covariance. The *income* is estimated to have a U-shape effect for the medical consultation, with a minimum between 800 AUD and 1150 AUD. The estimated effect of *gender* for Model A is slightly larger than for Model B.

Overall, the estimated effects of Model A are consistent with the results of Karlis and Ntzoufras.[30] In addition, we also considered a non-linear model. Both the linear and non-linear models indicated that the expected numbers of doctor consultations and prescribed medications increase with *age*. For *income*, the expected numbers of doctor visits and prescribed medications decreases with increasing *income* for Model A. The expected number of doctor visits also decreases in Model B as *income* increased, whereby a U-shaped effect for *income* can be observed.

Furthermore, because of the bivariate modeling, we also obtain information about the relationship between the outcomes. Here, both models showed a higher association between the number of doctor consultations and prescribed medications for women. Furthermore, Model B also included *age* and *income* as covariates for the covariance parameter and the model suggested that the association becomes greater with increasing *age*.

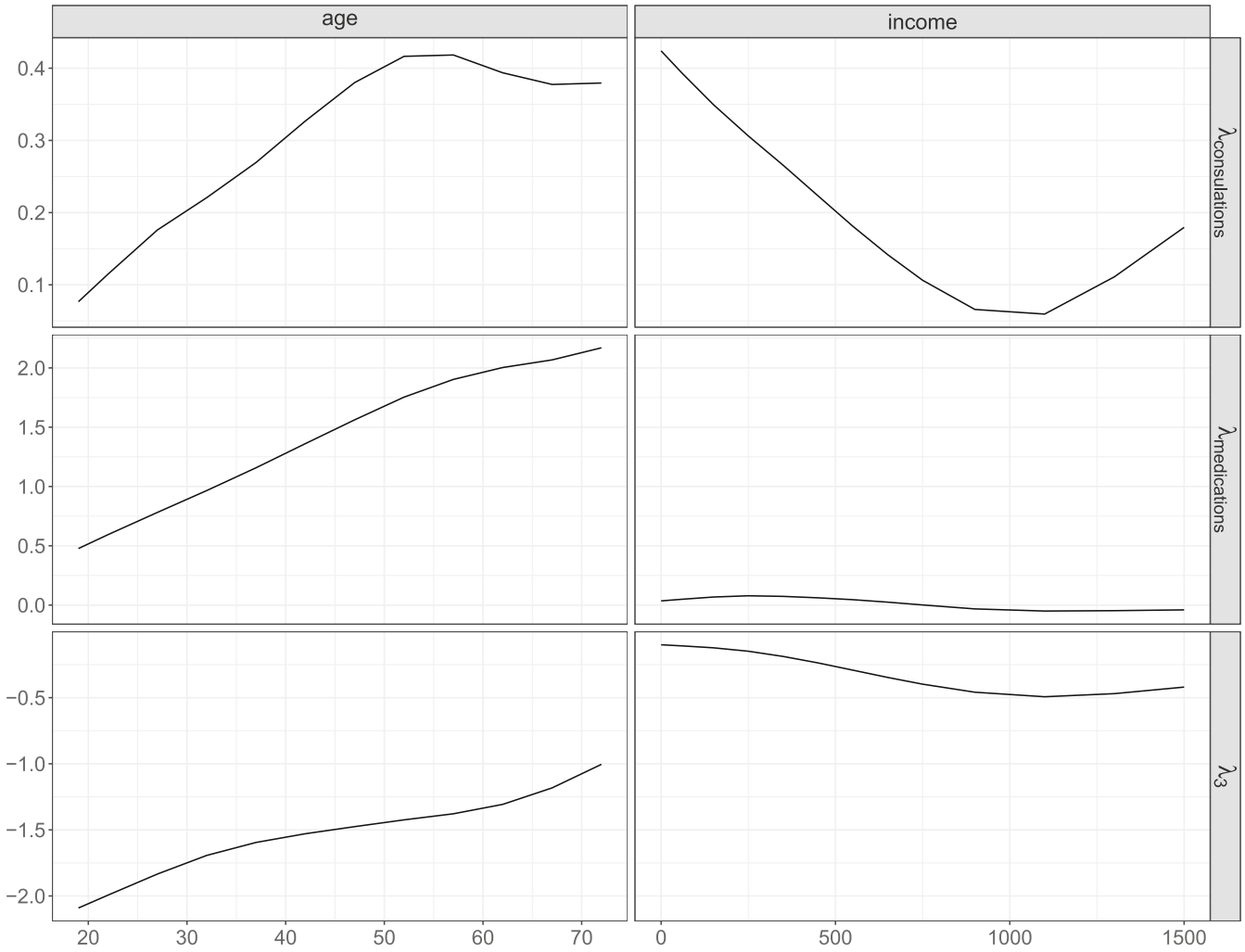

FIGURE 7 Partial effects of age and income on the demand for health care in Australia for model B

4.3 | Risk factors for undernutrition in Nigeria

To analyze childhood undernutrition, a large database is available from the Demographic and Health Survey (DHS, https://dhsprogram.com/), containing nationally representative information about the population's health and nutrition status in numerous developing and transition countries. Here, we consider a data set used in Klein et al.[64] which contains data from Nigeria collected in 2013 with overall 23,042 observations (after exclusion of outliers and inconsistent observations). The bivariate responses are *stunting*, which is defined as stunted growth measured as the insufficient height of the child concerning its age (chronic undernutrition), and *wasting*, which refers to insufficient weight for height (acute undernutrition). We analyze the joint distribution of these two responses using the bivariate Gaussian distribution with covariate-dependent marginal means and standard deviations as well as a covariate-dependent correlation parameter.

For continuous variables, P-splines were applied as base-learners, namely for *cage* (age of the child in months), *edupartner* (years of partner's education), *mage* (age of the mother in years) as well as *mbmi* (body mass index of the mother). Several other categorical covariates (12 covariates in total, eg, *bicycle*, *car*, *cbirthorder*) were included using simple linear models as base-learners. Furthermore, the neighborhood structure of the districts in Nigeria was incorporated and modeled by the spatial base-learner using a Markov random field. For a full description of the explanatory variables, see Appendix B.2. The stopping iteration of both models was tuned by 25-fold bootstrap and the step-length was set to $\nu = 0.1$.

Figures 8 and 9 show the results for the non-linear and spatial effects for all parameters. The estimated linear effects are given in Appendix Table B2. *Stunting* is estimated to be more affected by variables describing children's living situation, particularly *ctwin* (child is a twin) and the birth order (*cbirthorder*). Following our model, with higher birth order, the *stunting* score decreases, with negative values indicating that the children's growth is below the expected growth of a child

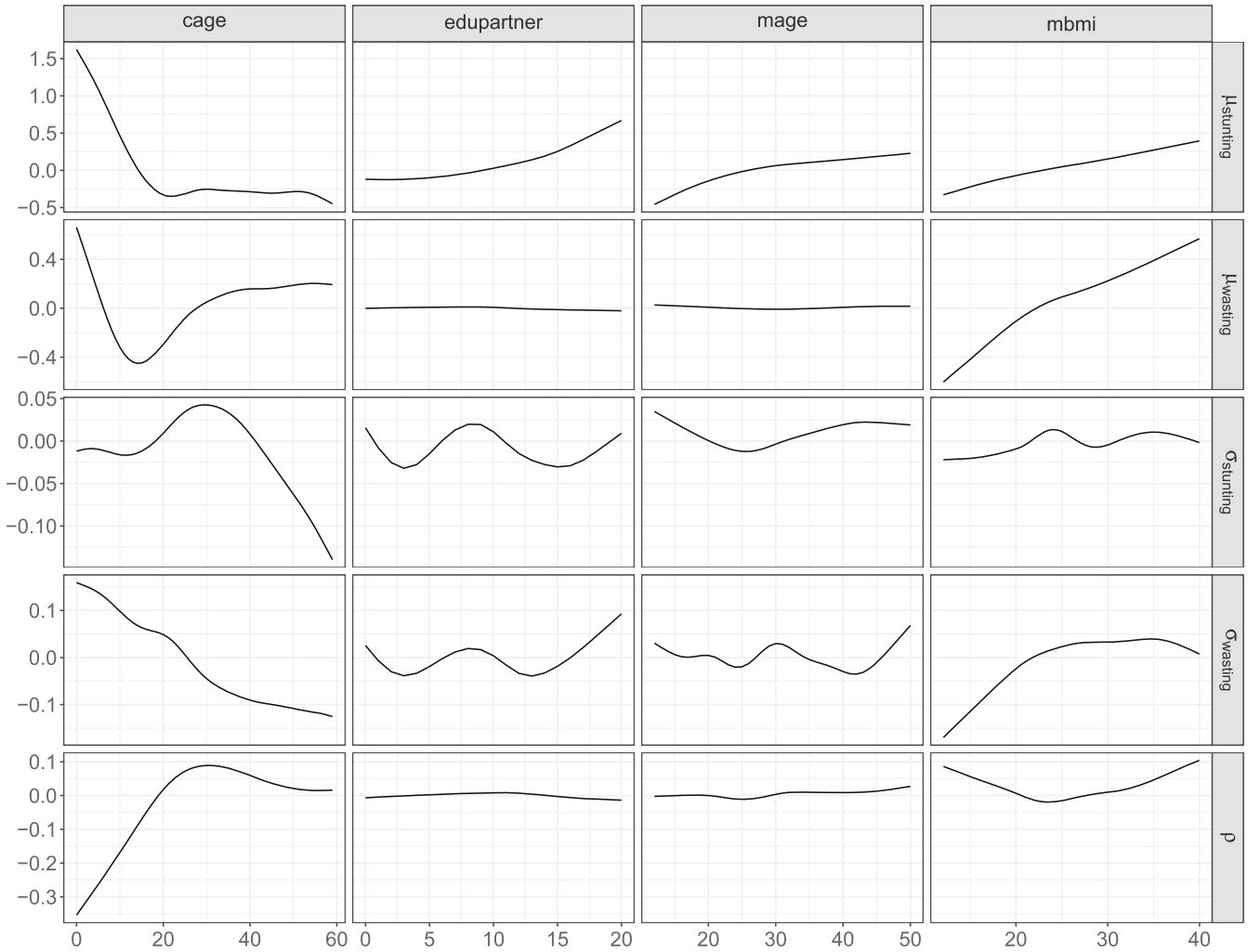

FIGURE 8 Non-linear effects of *cage*, *edupartner*, *mage* and *mbmi* for *stunting* and *wasting* of the bivariate Gaussian regression model for the Nigeria data

with normal nutrition. For wasting, *ctwin* had the largest effect, displaying also an increased risk for acute undernutrition. These results are in line with those of Klein et al.[64]

Furthermore, *stunting* and *wasting* were both influenced by *cage* and *mbmi* as well. Following our model, for *mbmi*, a higher BMI of the mother indicates a higher acute and chronic undernutrition. For *cage*, *stunting* and *wasting* is estimated to decrease (ie, risk increases) up to around 20 months. After 20 months, the risk for *wasting* is estimated to decrease again while remaining similar for *stunting*.

The scale parameter for *wasting*, for example, indicates a higher variability for children up to around 25 months. For children older than 25 months, the variability decreases slightly, whereby we observed a greater variability for stunting between 20 and 40 months. The correlation is negative for children younger than 20 months and is approximately zero after a small positive correlation between 20 and 50 months. This finding indicates an interaction between *stunting* and *wasting* depending on the child's age, which is non-linear and stronger for younger children. Thus, children with a greater height in the first years of life have a lower weight for height and vice versa. The other covariates have only a minor estimated effect on the correlation parameter. These results are consistent with previous findings,[11,64] which also holds for the spatial effects. The regional effect was selected to be informative for all distribution parameters. The effect of chronic undernutrition, for example, showed a lower risk of stunted growth in regions in southern Nigeria due to a positive effect. These regions also have a lower variability of chronic undernutrition compared to the average regions in the center of the county. This means that in this part of Nigeria the score for stunting is estimated to be on average lower and its variability is also smaller. By contrast, the regions in the north are estimated to have a higher risk for *stunting*. In terms of the correlation, some regions in the north are estimated to have a negative effect, while other regions in the south are

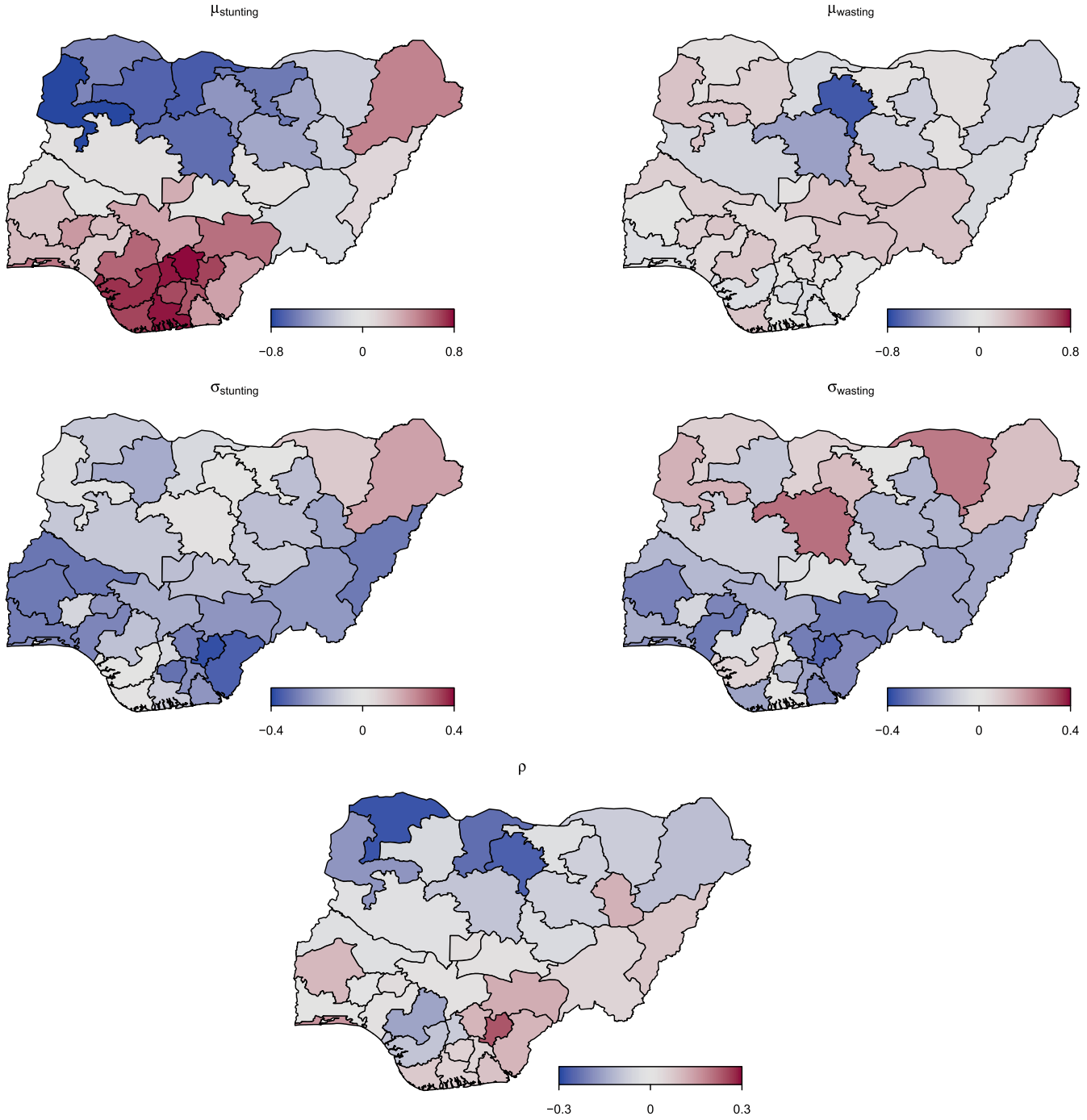

FIGURE 9 Spatial structure of *stunting* and *wasting* in Nigeria for the distribution parameters of the Gaussian distribution

estimated to show a slight positive effect on the correlation. A positive effect suggests that these regions have a problem of acute undernutrition as well as chronic undernutrition.

Overall, chronic undernutrition (*stunting*) is mostly affected by the living conditions of the children, for example, the birth order. Whereby, *stunting* and *wasting* were both influenced by the mother's BMI and particularly by the child's age. Additional effects of the covariates on the scale and correlation parameters also suggested greater uncertainty for younger children for acute undernutrition through a positive effect on the standard deviation, with variability decreasing with age. Furthermore, we observed a stronger negative correlation between *stunting* and *wasting* in younger children, that is, as stunting increases, wasting is expected to be lower. This means that children with a greater height tend to suffer from a lower weight for height at a younger age.

5 | DISCUSSION

We developed statistical boosting for modeling distributional regression with multivariate outcomes. Motivated by our biomedical applications, we considered three important multivariate parametric distributions: the bivariate Bernoulli, Poisson and Gaussian distributions. As special merits over classical maximum likelihood or Bayesian approaches to multivariate GAMLSS, our boosting framework can directly be used for high-dimensional data problems ($p > n$), while allowing for a data-driven variable selection mechanism that allows for sparse models for all parameters of a multivariate distribution.

In simulation studies, we have illustrated that the proposed boosting approach is able to identify the correct predictors in different data situations, including low- and high-dimensional settings and incorporating different effect types such as spatial effects. A comparison with the boosted univariate models showed that the bivariate models yielded more accurate estimates for the true structure of the effects. The wide applicability of our approach is illustrated on three different biomedical data sets, where we extend previous studies and also confirm findings from the literature. Applying our approach to examine jointly the genetic predisposition for chronic ischemic heart disease and high cholesterol not only provides information on the dependency of these phenotypes on the genetic variants, but also allows to identify the variants that affect the association between both phenotypes. This is in strong contrast to classical methods to estimate, for example, polygenic risk scores via accumulating effects from univariate linear models with single variants as predictor variables.[65] Our approach does not only incorporate multivariable predictor models, but also considers multivariate outcomes and hence allows to assess also the genetic predisposition for the association between several phenotypes, such as heart disease and high cholesterol. To the best of our knowledge, this is the first time multivariate distributional regression was adapted to model the joint genetic liability for multiple phenotypes.

In examining possible effects of patients characteristics on demand for health care, we found that age and income are relevant predictors, but also that gender affected the association between the number of doctor consultations and prescribed medications, with a stronger association found for women (cf. Karlis and Ntzoufras[30]).

In the third application analyzing the risk of undernutrition in Nigeria, an association was found between chronic undernutrition and the child's living condition. In addition, the age of the child had a relevant influence on all distribution parameters related to chronic and acute undernutrition; furthermore, the regional effect was selected not only for the margins but also for the scale and correlation parameters.

To summarize, the application of our approach to boost multivariate distributional regression is particularly beneficial in settings where at least some of the following criteria are met: (i) multiple associated responses are of interest; (ii) the association or other characteristics of the joint distribution depend on covariates; (iii) there are multiple explanatory variables available without clear prior-knowledge; (iv) the aim of the analysis is exploratory, hypothesis-generating or prediction.

A limitation regarding the considered distributions in our approach is the restriction of the Poisson distribution to positive dependency between the two responses. A possible solution for this restriction in future research could be the use of alternative parameterization,[39] which also allow for modeling negative correlations; however, these have the disadvantage that the interpretation of effects on these parameters becomes much more difficult. A limitation of our algorithm is the relatively high selection rates for variables with only minor importance, which occurs particularly in low-dimensional settings. In this context, Strömer et al.[66] have recently proposed an approach to deselect predictors with negligible impact to obtain sparser models with statistical boosting. We want to investigate the incorporation of this proposal in the context of multivariate GAMLSS in the future. Moreover, as the number of distribution parameters and the complexity of the model increases (eg, due to many non-linear effects), the algorithm becomes computationally more intensive. To address this problem, also alternative approaches for early stopping could be considered. A promising approach in the future which has been developed for univariate location models is probing, where randomly shuffled versions of the original observed variables (probes) are added to the data set and the algorithm stopped when the first probe is selected.[67] Furthermore, a unique fixed step-length for all distribution parameters (as it is currently good practice in statistical boosting) could lead in some settings to an imbalance in the updates of predictors. In the most extreme case, the algorithm stops before some of the distribution parameters received an update at all. This could be tackled by scaling the gradient vectors or directly the outcome variable.[50] Zhang et al.[48] recently proposed an approach for adaptive step-lengths in Gaussian location and scale models to find the optimal step-length for the parameters. Further research is warranted on extending adaptive step-lengths approaches beyond Gaussian models to the more complex multivariate regression models.

Last, our focus has been on bivariate distributional regression models, but we will consider extending the models to higher dimensional responses in future research. From an algorithmic perspective, the extension should be

straight-forward as it only adds more distribution parameters in our proposed framework. However, not only the construction of appropriate response distributions but also the interpretation of the effect estimates becomes more challenging. For example, for the multivariate Gaussian distribution, the main challenge is the parameterization of the covariance matrix and a promising route here could be based on a (modified) Cholesky decomposition.[68] Similar, the extension of the bivariate Poisson distribution to higher dimensions has some difficulties due to the complicated form of the joint probability function. The most common extension would force all the pairs of variables to have the same covariance,[69] whereby Karlis and Meligkotsidou[70] already discussed a model with a two-way covariance term that allows for different covariances between the variables.

ACKNOWLEDGEMENTS

The work on this article was supported by the Deutsche Forschungsgemeinschaft (DFG, grant number 428239776, KL3037/2-1, MA7304/1-1). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The code used for the simulations and the biomedical applications is available at GitHub https://github.com/AnnikaStr/DistRegBoost. The genomic cohort data is available upon request from the UK Biobank at https://www.ukbiobank.ac.uk/. This research has been conducted using the UK Biobank resource under application number 81202. The data set for health care in Australia is openly available from Cameron and Trivedi at http://cameron.econ.ucdavis.edu/racd/count.html. The Nigeria data set may be accessed upon request from the Demographic and Health Survey (DHS, https://dhsprogram.com/).

ORCID

Annika Strömer https://orcid.org/0000-0002-1284-3318
Nadja Klein https://orcid.org/0000-0001-5196-3374
Christian Staerk https://orcid.org/0000-0003-0526-0189
Hannah Klinkhammer https://orcid.org/0000-0003-3752-1275
Andreas Mayr https://orcid.org/0000-0001-7106-9732

REFERENCES

1. Hastie T, Tibshirani R. *Generalized Additive Models*. 1st ed. London: Chapman & Hall; 1990.
2. Wood SN. *Generalized Additive Models: An Introduction with R*. 2nd ed. London: Chapman & Hall/CRC; 2006.
3. Koenker R. *Quantile Regression*. 1st ed. United Kingdom: Cambridge University Press; 2005.
4. Offen W, Chuang-Stein C, Dmitrienko A, et al. Multiple co-primary endpoints: medical and statistical solutions: a report from the multiple endpoints expert team of the pharmaceutical research and manufacturers of America. *Drug Inf J*. 2007;41(1):31-46.
5. Zellner A. An efficient method of estimating seemingly unrelated regressions and test of aggregation bias. *J Am Stat Assoc*. 1962;57(298):348-368.
6. Lang S, Adebayo SB, Fahrmeir L, Steiner WJ. Bayesian geoadditive seemingly unrelated regression. *Comput Stat*. 2003;18:263-292.
7. King G. A seemingly unrelated Poisson regression model. *Sociol Methods Res*. 1989;17(3):235-255.
8. Gallant A. Seemingly unrelated nonlinear regressions. *J Econom*. 1975;3(1):35-50.
9. Fiebig DG. *Seemingly Unrelated Regression*. United Kingdom Oxford: Blackwell Publishers; 2001:101-121.
10. Rigby RA, Stasinopoulos DM. Generalized additive models for location, scale and shape. *J R Stat Soc: Ser C (Appl Stat)*. 2005;54(3):507-554.
11. Klein N, Kneib T, Klasen S, Lang S. Bayesian structured additive distributional regression for multivariate Responses. *J R Stat Soc. Ser C: Appl Stat*. 2015;64(4):569-591.
12. Cole TJ. Commentary: Methods for calculating growth trajectories and constructing growth centiles. *Stat Med*. 2019;38(19):3571-3579.
13. Papageorghiou AT, Ohuma EO, Altman DG, et al. International standards for fetal growth based on serial ultrasound measurements: the Fetal Growth Longitudinal Study of the INTERGROWTH-21st Project. *Lancet*. 2014;384(9946):869-879.
14. World Health Organization. *WHO Child Growth Standards: Length/Height for Age, Weight-for-Age, Weight-for-Length, Weight-for-Height and Body Mass Index-for-Age, Methods and Development*. Geneva, Switzerland: World Health Organization; 2006.
15. Hans N, Klein N, Faschingbauer F, Schneider M, Mayr A. Boosting distributional copula regression. *Biometrics*. 2022:1-13.
16. Zhu B, Dunson DB, Ashley-Koch AE. Adverse subpopulation regression for multivariate outcomes with high-dimensional predictors. *Stat Med*. 2012;31(29):4102-4113.
17. Wu C, Cui Y, Ma S. Integrative analysis of gene–environment interactions under a multi-response partially linear varying coefficient model. *Stat Med*. 2014;33(28):4988-4998.
18. Liu H, Sunil RJ. Generalized finite mixture of multivariate regressions with applications to therapeutic biomarker identification. *Stat Med*. 2020;39(28):4301-4324.

19. Friedman J, Hastie T, Tibshirani R. Additive logistic regression: a statistical view of boosting (With discussion and a rejoinder by the authors). *Ann Stat*. 2000;28(2):337-407.
20. Friedman J. Greedy function approximation: a gradient boosting machine. *Ann Stat*. 2001;29(5):1189-1232.
21. Bühlmann P, Hothorn T. Boosting algorithms: regularization, prediction and model fitting. *Stat Sci*. 2007;22(4):477-505.
22. Mayr A, Fenske N, Hofner B, Kneib T, Schmid M. 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(3):403-427.
23. Thomas J, Mayr A, Bischl B, Schmid M, Smith A, Hofner B. Gradient boosting for distributional regression: faster tuning and improved variable selection via noncyclical updates. *Stat Comput*. 2018;28:673-687.
24. Marshall AW, Olkin I. A family of bivariate distributions generated by the bivariate Bernoulli distribution. *J Am Stat Assoc*. 1985;80(390):332-338.
25. Kocherlakota S, Kocherlakota K. *Bivariate Discrete Distributions*. 1st ed. New York: Dekker; 1992.
26. Kotz S, Balakrishnan N, Johnson N. *Continuous Multivariate Distributions, Volume 1. Models and Applications,* 2nd ed. New York: John Wiley & sons; 2000.
27. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med*. 2015;12(3):e1001779.
28. Burgess S, Thompson S. *Mendelian Randomization: Methods for Using Genetic Variants in Causal Estimation*. 1st ed. New York: Chapman and Hall/CRC; 2015.
29. Cameron A, Trivedi P. *Regression Analysis of Count Data*. 1st ed. Cambridge, UK: Cambridge University Press; 1998.
30. Karlis D, Ntzoufras I. Bivariate Poisson and diagonal inflated bivariate Poisson regression models in R. *J Stat Softw*. 2005;14(10):1-36.
31. Fahrmeir L, Kneib T, Lang S. Penalized structured additive regression for space-time data: a Bayesian perspective. *Stat Sin*. 2004;14(3):731-761.
32. Fahrmeir L, Kneib T, Lang S, Marx B. *Regression: Models, Methods and Applications*. 1st ed. Berlin: Springer-Verlag; 2013.
33. Eilers PHC, Marx BD. Flexible smoothing with B-splines and penalties. *Stat Sci*. 1996;11(2):89-121.
34. Rue H, Held L. *Gaussian Markov Random Fields*. 1st ed. New York/Boca Raton: Chapman & Hall/CRC; 2005.
35. Johnson N, Kotz S, Balakrishnan N. *Discrete Multivariate Distributions*. 1st ed. New York: Wiley; 1997.
36. McCullagh P, Nelder J. *Generalized Linear Models*. Monographs on Statistics and Applied Probability Series. 2nd ed. London: Chapman and Hall/CRC; 1989.
37. Palmgren J. Regression models for bivariate binary responses. *UW Biostatistics Working Paper Series*. 1989 Working Paper 101.
38. Dale JR. Global cross-ratio models for bivariate, discrete, ordered Responses. *Biometrics*. 1986;42(4):909-917.
39. Lakshminarayana J, Pandit S, Rao KS. On a bivariate Poisson distribution. *Commun Stat Theory Methods*. 1999;28(2):267-276.
40. Ma Z, Hanson TE, Ho YY. Flexible bivariate correlated count data regression. *Stat Med*. 2020;39(25):3476-3490.
41. Freund Y. Boosting a weak learning algorithm by majority. *Inf Comput*. 1995;12(2):256-285.
42. Li Z, Luo Z, Sun Y. Robust nonparametric integrative analysis to decipher heterogeneity and commonality across subgroups using sparse boosting. *Stat Med*. 2022;41(9):1658-1687.
43. Wu M, Ma S. Robust semiparametric gene-environment interaction analysis using sparse boosting. *Stat Med*. 2019;38(23):4625-4641.
44. Tutz G, Binder H. Generalized additive modeling with implicit variable selection by likelihood-based boosting. *Biometrics*. 2006;62(4):961-971.
45. Mayr A, Binder H, Gefeller O, Schmid M. The evolution of boosting algorithms. From machine learning to statistical modelling. *Methods Inf Med*. 2014;53(6):419-427.
46. Mayr A, Binder H, Gefeller O, Schmid M. Extending statistical boosting. An overview of recent methodological developments. *Methods Inf Med*. 2014;53(6):428-435.
47. Hofner B, Mayr A, Robinzonov N, Schmid M. Model-based boosting in R: a hands-on tutorial using the R package mboost. *Comput Stat*. 2014;29:3-35.
48. Zhang B, Hepp T, Greven S, Bergherr E. Adaptive step-length selection in gradient boosting for Gaussian location and scale models. *Comput Stat*. 2022;37:2295-2332.
49. Mayr A, Hofner B, Schmid M. The importance of knowing when to stop. A sequential stopping rule for component-wise gradient boosting. *Methods Inf Med*. 2012;51(2):178-186.
50. Hofner B, Mayr A, Schmid M. gamboostLSS: an R package for model building and variable selection in the GAMLSS framework. *J Stat Softw*. 2016;74(1):1-31.
51. Gneiting T, Stanberry LI, Grimit EP, Held L, Johnson NA. Assessing probabilistic forecasts of multivariate quantities, with an application to ensemble predictions of surface winds. *TEST*. 2008;17:211-235.
52. Brier GW. Verification of forecasts expressed in terms of probability. *Mon Weather Rev*. 1950;78(1):1-3.
53. Yee TW. The VGAM package for categorical data analysis. *J Stat Softw*. 2010;32(10):1-34.
54. Wolodzko T. extraDistr: additional univariate and multivariate distributions. R package version 1.9.1. 2020.
55. Umlauf N, Klein N, Simon T, Zeileis A. bamlss: a Lego toolbox for flexible Bayesian regression (and beyond). *J Stat Softw*. 2021;100:1-53. doi:10.18637/jss.v100.i04
56. Linsel-Nitschke P, Götz A, Erdmann J, et al. Lifelong reduction of LDL-cholesterol related to a common variant in the LDL-receptor gene decreases the risk of coronary artery disease – A Mendelian randomisation study. *PLOS One*. 2008;3(8):1-9.
57. 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(3):1-22.

58. Purcell S. Chang C. Plink 2.0. 2015 https://www.cog-genomics.org/plink/2.0/
59. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. Second-generation PLINK: rising to the challenge of larger and richer datasets. *Giga Sci*. 2015;4(1):1-16.
60. Privé F, Aschard H, Ziyatdinov A, Blum MGB. Efficient analysis of large-scale genome-wide data with two R packages: bigstatsr and bigsnpr. *Bioinform*. 2018;34(16):2781-2787.
61. Maj C, Staerk C, Borisov O, et al. Statistical learning for sparser fine-mapped polygenic models: the prediction of LDL-cholesterol. *Genet Epidemiol*. 2022;46(8):589-603.
62. Klinkhammer H, Staerk C, Maj C, Krawitz PM, Mayr A. A statistical boosting framework for polygenic risk scores based on large-scale genotype data. *Front Genet*. 2023;13:1-16. doi:10.3389/fgene.2022.1076440
63. Elosua R, Sayols-Baixeras S. The genetics of ischemic heart disease: from current knowledge to clinical implications. *Revista Española de Cardiología (English Edition)*. 2017;70(9):754-762.
64. Klein N, Carlan M, Kneib T, Lang S, Wagner H. Bayesian effect selection in structured additive distributional regression models. *Bayesian Anal*. 2021;16(2):545-573.
65. Choi SW, Mak TSH, O'Reilly PF. Tutorial: a guide to performing polygenic risk score analyses. *Nat Protoc*. 2020;15:2759-2772.
66. Strömer A, Staerk C, Klein N, Weinhold L, Titze S, Mayr A. Deselection of base-learners for statistical boosting – with an application to distributional regression. *Stat Methods Med Res*. 2022;31(2):207-224.
67. Thomas J, Hepp T, Mayr A, Bischl B. Probing for sparse and fast variable selection with model-based boosting. *Comput Math Methods Med*. 2017.
68. Pourahmadi M. Covariance estimation: the GLM and regularization perspectives. *Stat Sci*. 2011;26(3):369-387.
69. Karlis D. An EM algorithm for multivariate poisson distribution and related models. *J Appl Stat*. 2003;30(1):63-77.
70. Karlis D, Meligkotsidou L. Multivariate poisson regression with covariance structure. *Stat Comput*. 2005;15:255-265.

How to cite this article: Strömer A, Klein N, Staerk C, Klinkhammer H, Mayr A. Boosting multivariate structured additive distributional regression models. *Statistics in Medicine*. 2023;42(11):1779-1801. doi: 10.1002/sim.9699