ROYAL STATISTICAL SOCIETY
DATA | EVIDENCE | DECISIONS

Journal of the Royal Statistical Society
Applied Statistics
Series C

Appl. Statist. (2015)
64, *Part* 4, *pp.* 569–591

Bayesian structured additive distributional regression for multivariate responses

Nadja Klein, Thomas Kneib and Stephan Klasen

Georg-August-Universität Göttingen, Germany

and Stefan Lang

University of Innsbruck, Austria

[Received November 2013. Revised August 2014]

Summary. We propose a unified Bayesian approach for multivariate structured additive distributional regression analysis comprising a huge class of continuous, discrete and latent multivariate response distributions, where each parameter of these potentially complex distributions is modelled by a structured additive predictor. The latter is an additive composition of different types of covariate effects, e.g. non-linear effects of continuous covariates, random effects, spatial effects or interaction effects. Inference is realized by a generic, computationally efficient Markov chain Monte Carlo algorithm based on iteratively weighted least squares approximations and with multivariate Gaussian priors to enforce specific properties of functional effects. Applications to illustrate our approach include a joint model of risk factors for chronic and acute childhood undernutrition in India and ecological regressions studying the drivers of election results in Germany.

Keywords: Correlated responses; Dirichlet regression; Iteratively weighted least squares proposal; Markov chain Monte Carlo simulation; Penalized splines; Seemingly unrelated regression

1. Introduction

Whereas most regression models are formulated for one single response variable alone, analysing multivariate, correlated response variables simultaneously can be of considerable relevance. Thereby the nature of the responses can be rather different, ranging from discrete multicategorical to multivariate continuous data types. Although multicategorical data have been treated extensively in the literature (mostly within a Bayesian framework; see for example Chen and Dey (2000), Albert and Chib (1993), Imai and van Dyk (2005) and Frühwirth-Schnatter *et al.* (2009)), less emphasis has been placed on the situation of multivariate continuous responses, despite applied interest in this problem. In this paper, we shall consider two case-studies representing different types of multivariate continuous responses. In the first case-study, which is presented in detail in Section 3, we deal with the simultaneous analysis of two indicators for undernutrition of children in India (as a representative case of a developing country). The two indicators measure chronic and acute forms of undernutrition and therefore represent two different yet related outcomes where it is of particular interest to study the effect of covariates on the correlation between the two. For example, it will be of interest to investigate how the cor-

Address for correspondence: Nadja Klein, Georg-August-Universität Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany.
E-mail: nklein@uni-goettingen.de

 0035–9254/15/64569

relation changes with the age of the child. In our second case-study (see Section 4), we analyse socio-economic determinants of election results for Germany's Federal Parliament in 2009. The multivariate response vector then represents shares of votes for different parties achieved in the election and correlations are generated by the constraint that shares need to sum to 1.

So far, most publications on regression analyses for multivariate responses have focused on one specific distribution such as the multivariate normal distribution in seemingly unrelated regression models (Zellner, 1962; Greene, 2011) or distributions for categorical outcomes and count data; see for example Winkelmann (2008) or Tutz (2011) for recent overviews. In addition, most of these approaches have focused exclusively on linear predictors following the classical framework of generalized linear models (McCullagh and Nelder, 1989; Fahrmeir *et al.*, 2013). However, the restriction to a parametric predictor does not capture the flexibility of modelling (possibly more realistic) non-linear effects of covariates or spatial variation within the data. Only a few references are available dealing with multivariate responses and non-parametric predictors. For instance, Fahrmeir and Lang (2001) proposed multicategorical regression models in the spirit of generalized additive models (Hastie and Tibshirani, 1990; Ruppert *et al.*, 2003; Wood, 2006; Fahrmeir *et al.*, 2013) from a Bayesian point of view. Bayesian seemingly unrelated regression models with semiparametric predictors have been developed in Lang *et al.* (2003). However, in all these models only the mean of the components of the response is related to covariates, neglecting the potential dependence of higher moments or correlations of the response vector on covariates. Although covariate effects are indeed straightforward to estimate and easy to interpret in mean regression models, this simplified modelling approach may lead to model misspecification and therefore invalid inferential conclusions. Smith and Kohn (2000) for example showed in simulation studies that estimates can become inefficient and that non-linear effects can be biased when applying univariate regressions instead of a multivariate model. Accordingly, it is of great interest to provide a framework that is sufficiently flexible to model more than just the mean while remaining interpretable and reliable. The aforementioned problems can be solved by the framework of generalized additive models for location, scale and shape (Rigby and Stasinopoulos, 2005) where potentially complex parametric distributions are assumed for the response variable. Additionally, each parameter of the distribution, i.e. variances and further moments, can be modelled in terms of covariates since they are related to additive regression predictors. Estimates for a large number of different types of distribution are obtained from Newton–Raphson or Fisher scoring types of algorithm used to maximize the (penalized) likelihood. However, the framework of generalized additive models for location, scale and shape is currently restricted to univariate responses.

To address this gap, we extend Bayesian distributional regression for univariate responses, recently proposed by Klein *et al.* (2013), to a generic approach for multivariate responses in the spirit of generalized additive models for location, scale and shape. The notion of distributional regression is more general than generalized additive models for location, scale and shape since the parameters of the response distributions do not always relate directly to location, scale or shape but instead functions of several parameters usually lead to these characteristics. Inference will be realized by a Markov chain Monte Carlo (MCMC) simulation algorithm based on distribution-specific iteratively weighted least squares approximations to the full conditionals. The approach is implemented in the free software package BayesX (`www.bayesx.org`) and will be available with version 3.0. Code for reproducing our analyses is available from

`http://wileyonlinelibrary.com/journal/rss-datasets`

Concerning the choice of the multivariate response distribution, we consider the following special cases. The normal distribution is one of a few exceptions where the parameters of the

response distribution are directly interpretable since they represent expectation and variance of the response. This favourable property is preserved even in the bivariate case where, in addition to the first and second moment of the components of the response vector, the correlation parameter can be explained by various covariate effects (although the situation becomes more difficult when going beyond two dimensions; see the discussion below). We shall use the bivariate normal distribution in our application on childhood undernutrition in India in Section 3 to develop a joint model for chronic and acute undernutrition. We also consider an extension based on the bivariate t-distribution to contrast the bivariate normal distribution with an alternative with heavier tails. The performance of Bayesian inference in bivariate normal and bivariate t-models is also evaluated in two simulation studies (see the on-line supplement sections C and D and the summary at the end of Section 2.3). In Appendix A, we present an extension of the multivariate normal model to the multivariate probit model as an example of binary multivariate regression which is often employed in economic and biostatistical research.

In our second application on German elections, the response vector consists of fractions of electoral votes for five different parties and one category summarizing the remaining votes for smaller parties. Hence, a restriction is given by the fact that the sum of all proportions equals 1 and the positive density can be represented by a five-dimensional open simplex. A natural candidate for analysing such fractions is therefore the Dirichlet distribution with six positive parameters where none of these parameters is directly linked to location, scale or shape but ratios or the sum of several parameters can be interpreted more meaningfully.

The rest of the paper is structured as follows: in Section 2, we first introduce different multivariate regression models in more detail and present a generic formulation for inference in Bayesian structured additive distributional regression for multivariate responses. Section 2 also comments on model choice in multivariate distributional regression. Section 3 illustrates the analysis of the Indian undernutrition data with bivariate geoadditive regression models whereas Section 4 contains results of the ecological regression for the election data. The final Section 5 concludes and provides comments on directions of future research.

2. Bayesian multivariate distributional regression

2.1. Observation models

Let $p(y_{i1}, \ldots, y_{iD}|\vartheta_{i1}, \ldots, \vartheta_{iK}) \equiv p_i$, $i = 1, \ldots, n$, be the conditional K-parametric densities of D-dimensional random variables $(y_{i1}, \ldots, y_{iD})'$ given some covariate information $\boldsymbol{\nu}_i$. The basic idea of distributional regression is to link each parameter ϑ_{ik} to a semiparametric structured additive predictor $\eta_i^{\vartheta_k}$ formed of the covariates with the help of monotone, twice differentiable response functions h_k, such that $\vartheta_{ik} = h_k(\eta_i^{\vartheta_k})$ and $\eta_i^{\vartheta_k} = h_k^{-1}(\vartheta_{ik})$. Note that each predictor and each distribution parameter can be summarized in vectors of length n, $\boldsymbol{\eta}^{\vartheta_k} = (\eta_1^{\vartheta_k}, \ldots, \eta_n^{\vartheta_k})'$ and $\boldsymbol{\vartheta}_k = (\vartheta_{1k}, \ldots \vartheta_{nk})'$, and that the superscript ϑ_k in the predictor is a notation to indicate which parameter the predictor belongs to. The response function is usually chosen to maintain restrictions on the parameter space, like the exponential function $\vartheta_{ik} = \exp(\eta_i^{\vartheta_k})$ to ensure a parameter with values on the positive real half-axis, the identity function if the parameter space is unrestricted or $\vartheta_{ik} = \eta_i^{\vartheta_k}/\sqrt{\{1 + (\eta_i^{\vartheta_k})^2\}}$ if $\vartheta_{ik} \in [-1, 1]$, which is for example the case for the correlation between two variables.

2.1.1. Examples of multivariate response distributions

In what follows, we describe examples of multivariate distributions that play important and useful roles in applied research in general and in particular in the case-studies that are pre-

sented in Sections 3 and 4. Note, however, that more parametric distributions may be added by transferring the inferential procedure that is introduced in Section 2.3.

2.1.1.1. Multivariate continuous distributions. For the simultaneous analysis of multiple continuous response variables without restrictions on the domain, we require appropriate distributions including the possibility of correlations between the responses. The most prominent example is provided by the multivariate normal distribution where for a D-dimensional random vector $\mathbf{y} = (y_1, \ldots, y_D)'$ we write $\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with expectation $\boldsymbol{\mu} = (\mathbb{E}(y_1), \ldots, \mathbb{E}(y_D))' \in \mathbb{R}^D$ and a positive semidefinite covariance matrix $\boldsymbol{\Sigma} = \mathrm{cov}(y_i, y_j) \in \mathbb{R}^{D \times D}$ for $i, j = 1, \ldots, D$. We assume that $\boldsymbol{\Sigma}$ is strictly positive definite, so that the density of $\mathbf{y}$ is defined as

$$p(y_1, \ldots, y_D) = \frac{1}{\sqrt{\{(2\pi)^D \det(\boldsymbol{\Sigma})\}}} \exp\left\{ -\frac{1}{2} (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu}) \right\}.$$

The multivariate normal distribution has several practically and theoretically attractive properties; see for example Kotz *et al.* (2005). However, finding a statistically interpretable and unconstrained parameterization for the covariance matrix is still a challenge when the dimension D is larger than 2. Possible approaches would be to work with the variance–correlation decomposition, the spectral decomposition or the (modified) Cholesky decomposition. The first approach has the appeal that factors can easily be interpreted in terms of standard deviations and correlations but is inconvenient in regression models with $D > 2$ because complex constraints are required to ensure positive definiteness of the resulting covariance matrix. The second approach results in parameters that can be interpreted as variances and coefficients of principal components but the latter also induce constraints due to their orthogonality. Finally, with the Cholesky decomposition, the interpretation of estimated effects becomes difficult. However, in a modified decomposition into two triangular and one diagonal matrix $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{D}^2\mathbf{L}'$, Pourahmadi (2011) showed that the non-redundant entries of $\mathbf{L}^{-1}$ are unconstrained and statistically meaningful such that separate regression specifications can be formulated for each of the lower triangular elements of $\mathbf{L}^{-1}$ and the diagonal elements of $\mathbf{D}$.

For the remainder of this paper, we restrict ourselves to the case $D = 2$ where the variance–correlation decomposition can be easily applied and $\boldsymbol{\Sigma}$ becomes

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

with $\sigma_1^2 = \mathrm{var}(y_1)$, $\sigma_2^2 = \mathrm{var}(y_2)$ and $\rho = \mathrm{corr}(y_1, y_2)$. In distributional regression, the expectations and standard deviations of the marginal distributions as well as the correlation parameter can be estimated as functions of covariates by applying the transformations

$$\begin{aligned} \eta_i^{\mu_1} &= \mu_{i1}, & \eta_i^{\mu_2} &= \mu_{i2}, \\ \eta_i^{\sigma_1} &= \log(\sigma_{i1}), & \eta_i^{\sigma_2} &= \log(\sigma_{i2}), \\ & & \eta_i^{\rho} &= \rho_i / \sqrt{(1 - \rho_i^2)}. \end{aligned}$$

We shall apply the bivariate normal distribution in our application to childhood undernutrition in India (Section 3) and compare it with the bivariate t-distribution, which is an alternative to the bivariate normal distribution with heavier tails that in general is considered to be less sensitive with respect to extreme observations. The multivariate t-distribution (Kotz *et al.*, 2005) is a multidimensional generalization of the univariate t-distribution. A D-dimensional random variable $\mathbf{y} = (y_1, \ldots, y_D)'$ is said to follow a D-dimensional t-distribution, i.e. $\mathbf{y} \sim t(n_{\mathrm{df}}, \boldsymbol{\mu}, \boldsymbol{\Sigma})$

with location parameters $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)'$, dispersion matrix $\boldsymbol{\Sigma}$ and degrees of freedom $n_{\text{df}} > 0$ if the density of $\mathbf{y}$ is given by

$$p(y_1, \ldots, y_D) = \frac{\Gamma\{(n_{\text{df}} + 2)/2\}}{\Gamma(n_{\text{df}}/2)(n_{\text{df}}\pi)^{D/2}} \det(\boldsymbol{\Sigma})^{-1/2} \{1 + (\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\mathbf{y} - \boldsymbol{\mu})\}^{-(n_{\text{df}} + D)/2}.$$

Compared with multivariate normal regression we obtain an additional predictor

$$\eta_i^{n_{\text{df}}} = \log(n_{\text{df},i})$$

for the degrees of freedom.

2.1.1.2. Multivariate binary distribution. The multivariate probit model is a generalization of the univariate probit model which can be used to estimate several correlated binary outcomes jointly. This model is of particular interest to researchers since it allows for the estimation of the treatment effect that a binary endogenous variable has on a binary outcome in the presence of unobservables (Heckman, 1978; Maddala, 1983; Woolridge, 2002). From a Bayesian point of view, inference can be performed based on a latent model representation that allows estimation of a complex correlation structure on the components of the response: assume a D-variate probit model with dependent binary variable $\mathbf{y} = (y_1, \ldots, y_D)'$ and corresponding unobservable latent variable $\mathbf{y}^* = (y_1^*, \ldots, y_D^*)'$. Similarly to the univariate case we assume a multivariate normal distribution for $\mathbf{y}^*$ and write

$$\mathbf{y}^* = \boldsymbol{\eta}^{\mu} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \boldsymbol{\Sigma}),$$

with $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_D)' = (\mathbb{E}(y_1^*), \ldots, \mathbb{E}(y_D^*))'$ and $\boldsymbol{\Sigma}$ equals $\text{corr}(y_{d_1}^*, y_{d_2}^*)$ if $d_1 \neq d_2$ and 1 otherwise for $d_1, d_2 = 1, \ldots, D$. Then, $\mathbf{y}$ is an indicator for whether the latent variable $\mathbf{y}^*$ is positive, i.e.

$$y_{id} = 1 \Leftrightarrow y_{id}^* > 0, \qquad i = 1, \ldots, n, \quad d = 1, \ldots, D.$$

Following the ideas of Albert and Chib (1993) we show in Appendix A that Bayesian inference in the multivariate probit model can be realized with the same quantities as for the multivariate normal distribution such that no further computations are necessary apart from the imputation of the latent responses $\mathbf{y}^*$.

2.1.1.3. Dirichlet distribution. In our second application on election results, the response quantity of interest is given by a vector containing the shares of votes for different parties. Naturally, the sum of the elements of this vector equals 1 and this property should be preserved by a statistical model. Such a model is obtained by assuming that $\mathbf{y} = (y_1, \ldots, y_D)'$, $D \geqslant 2$, follows a Dirichlet distribution (i.e. $\mathbf{y} \sim \text{Dir}(\boldsymbol{\alpha})$) with parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_D) \in \mathbb{R}_{>0}^D$ and density

$$p(y_1, \ldots, y_{D-1}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{d=1}^{D} y_d^{\alpha_d - 1},$$

$$\sum_{d=1}^{D} y_d = 1, \qquad y_d \geqslant 0,$$

where the normalizing constant is the multinomial beta function of $\boldsymbol{\alpha}$:

$$B(\boldsymbol{\alpha}) = \frac{\prod_{d=1}^{D} \Gamma(\alpha_d)}{\Gamma\left(\sum_{d=1}^{D} \alpha_d\right)}.$$

Some important properties of the Dirichlet distribution are as follows.

(a) The Dirichlet distribution is a generalization of the beta distribution since in the cases of $D=2$ it can be seen easily that y_1 is beta distributed, i.e. $y_1 \sim \text{Be}(\alpha_1, \alpha_2)$. In addition, the univariate marginal distributions of all components y_d are beta distributions with parameters α_d and $-\alpha_d + \Sigma_{k=1}^{D} \alpha_k$.
(b) When all α_d are equal to 1, the Dirichlet distribution reduces to the uniform distribution on the simplex defined by $\Sigma_{d=1}^{D} y_d = 1$, $y_d > 0$, into $\mathbb{R}^D$.
(c) The density f is zero outside the open $(D-1)$-dimensional simplex and could be defined on any subset of size $D-1$ because $y_D = 1 - \Sigma_{d=1}^{D-1} y_d$.
(d) The value y_d can be interpreted as the probability that an event will fall in category d with expectation

$$\mathbb{E}(y_d) = \alpha_d / \alpha_0$$

where $\alpha_0 = \Sigma_{d=1}^{D} \alpha_d$ represents a precision parameter.

In our approach to Dirichlet regression, each parameter α_d is linked to a structured additive predictor, i.e. $\eta_i^{\alpha_d} = \log(\alpha_{id})$.

2.2. Generic regression formulation

For any parameter of the multivariate distributions that were discussed in the previous section, the semiparametric predictor has the general form

$$\eta_i^{\vartheta_k} = \sum_{j=1}^{J_k} f_j^{\vartheta_k}(\boldsymbol{\nu}_i) \tag{1}$$

comprising various functions $f_j^{\vartheta_k}(\boldsymbol{\nu}_i)$ defined on the complete covariate information $\boldsymbol{\nu}_i$. Specific components may for instance be given by

(a) linear functions $f_j^{\vartheta_k}(\boldsymbol{\nu}_i) = \mathbf{x}_i' \beta_j^{\vartheta_k}$, including the overall level of the predictor as an intercept $\beta_{0j}^{\vartheta_k}$ and $\mathbf{x}_i$ is a subvector of $\boldsymbol{\nu}_i$ ($\mathbf{x}_i$ may be chosen specifically for each parameter ϑ_k but we suppress this potential dependence in our notation),
(b) continuous functions $f_j^{\vartheta_k}(\boldsymbol{\nu}_i) = f_j^{\vartheta_k}(x_i)$, where x_i is a single element of $\boldsymbol{\nu}_i$ and f_j is an appropriate smooth function to represent the effect of x_i on ϑ_{ik},
(c) spatial variations $f_j^{\vartheta_k}(\boldsymbol{\nu}_i) = f_j^{\vartheta_k}(s_i)$, where s_i represents spatial information, e.g. coordinate information in terms of longitude and latitude or discrete spatial information representing a fixed set of (administrative) geographical units, and
(d) random effects $f_j^{\vartheta_k}(\boldsymbol{\nu}_i) = \beta_{j,g_i}^{\vartheta_k}$, where g_i is a cluster variable that groups the observations.

We represent smooth terms in the models by using basis functions, as a result of which the predictors can always be written in the generic matrix notation

$$\boldsymbol{\eta}^{\vartheta_k} = \sum_{j=1}^{J_k} \mathbf{Z}_j^{\vartheta_k} \boldsymbol{\beta}_j^{\vartheta_k}$$

where the design matrices $\mathbf{Z}_j^{\vartheta_k}$ are obtained by evaluating the basis functions at observed covariate values and $\boldsymbol{\beta}_j^{\vartheta_k}$ are the vectors of basis coefficients to be estimated. Specific properties of the

basis coefficients such as smoothness are regularized by assuming possibly improper Gaussian priors

$$p\{\boldsymbol{\beta}_j^{\vartheta_k}|(\tau_j^{\vartheta_k})^2\} \propto \left\{\frac{1}{(\tau_j^{\vartheta_k})^2}\right\}^{rk(\mathbf{K}_j^{\vartheta_k})/2} \exp\left\{-\frac{1}{2(\tau_j^{\vartheta_k})^2}(\boldsymbol{\beta}_j^{\vartheta_k})'\mathbf{K}_j^{\vartheta_k}\boldsymbol{\beta}_j^{\vartheta_k}\right\} \tag{2}$$

where $\mathbf{K}_j^{\vartheta_k}$ is a prior precision matrix and $(\tau_j^{\vartheta_k})^2$ are smoothing variances. The latter are supplemented with inverse gamma hyperpriors, i.e. $(\tau_j^{\vartheta_k})^2 \sim \text{IG}(a_j^{\vartheta_k}, b_j^{\vartheta_k})$, to obtain a data-driven amount of smoothness with $a_j^{\vartheta_k} = b_j^{\vartheta_k} = 0.001$ as default values for practical analyses.

As a result, each term $\mathbf{f}_j^{\vartheta_k} = (f_j^{\vartheta_k}(\boldsymbol{\nu}_1), \ldots, f_j^{\vartheta_k}(\boldsymbol{\nu}_n))' = \mathbf{Z}_j^{\vartheta_k}\boldsymbol{\beta}_j^{\vartheta_k}$ is determined by a design matrix $\mathbf{Z}_j^{\vartheta_k}$ and a prior precision or penalty matrix $\mathbf{K}_j^{\vartheta_k}$. We give specific examples in what follows (suppressing the index j and superscript ϑ_k).

2.2.1. Continuous covariates

To approximate potentially non-linear effects, we use Bayesian P-splines; see Eilers and Marx (1996) and Brezger and Lang (2006) for detailed explanations. The $n \times S$ design matrix $\mathbf{Z}$ in this setting is composed of S B-spline basis functions evaluated at observed covariates x_i. Assuming equidistant knots for the spline expansion, a first- or second-order random walk is a sensible choice for the prior of $\boldsymbol{\beta}$, i.e.

$$\beta_s|\beta_{s-1}, \tau^2 \sim N(\beta_{s-1}, \tau^2), \qquad s = 2, \ldots, S,$$

or

$$\beta_s|\beta_{s-1}, \beta_{s-2}, \tau^2 \sim N(2\beta_{s-1} - \beta_{s-2}, \tau^2), \qquad s = 3, \ldots, S,$$

with non-informative priors for initial values. This prior structure yields the penalty matrix $\mathbf{K} = \mathbf{D}'\mathbf{D}$ where $\mathbf{D}$ is a difference matrix of first or second order.

2.2.2. Spatial effects

For discrete spatial effects observed on a lattice or regions, we consider Markov random fields; see Rue and Held (2005). Let $s_i \in \{1, \ldots, S\}$ denote the index or region observation i belongs to. Then $f(s_i) = \beta_{s_i}$ is assumed such that we estimate separate parameters $\beta_1, \ldots, \beta_S$ for each region. As a consequence, the $n \times S$ design matrix is an incidence matrix, i.e. $\mathbf{Z}[i, s]$ equals 1 if observation i belongs to location s and 0 otherwise. The simplest Markov random-field prior for the coefficients β_s is defined by

$$\beta_s|\beta_r, r \neq s, \tau^2 \sim N\left(\sum_{r \in \partial_s} \frac{1}{N_s}\beta_r, \frac{\tau^2}{N_s}\right),$$

where ∂_s denotes the set of neighbours of region s and N_s is the number of regions in ∂_s. The penalty matrix is then given by

$$\mathbf{K}[s, r] = \begin{cases} -1 & s \neq r, \quad r \in \partial_s, \\ 0 & s \neq r, \quad r \notin \partial_s, \\ N_s & s = r. \end{cases}$$

For detailed explanations on structured additive regression with further examples we refer the reader to Fahrmeir *et al*. (2013).

2.3. Bayesian inference with a generic algorithm

Bayesian inference in multivariate distributional regression is facilitated by MCMC simulation

techniques that allow us to implement a convenient divide-and-conquer strategy to make the complex posterior specifications tractable. Depending on the response distribution, the full conditionals $\log\{p(\boldsymbol{\beta}_j^{\vartheta_k}|\cdot)\}$ for the coefficient vectors of several distribution parameters ϑ_k may not be written in a closed form. In contrast, in the bivariate normal distribution for instance, it can be shown that the Gaussian priors yield a conjugate model for the parameters μ_d, $d=1,2$, such that the full conditionals for the regression coefficients corresponding to the expectation parameters are again Gaussian. In what follows, we, however, describe the situation where the full conditionals are not analytically accessible and note that, in cases in which the full conditionals can be obtained explicitly, the resulting Metropolis–Hastings updates are reduced to Gibbs samplers from multivariate normal distributions with the same parameters that we shall propose in the approximations to the full conditionals in more complicated situations.

More precisely, a quadratic Taylor series expansion of the log-likelihood function around the mode leads to iteratively weighted least square proposals (Gamerman, 1997; Brezger and Lang, 2006), i.e. multivariate normal proposal densities

$$\boldsymbol{\beta}_j^{\vartheta_k} \sim N\{\boldsymbol{\mu}_j^{\vartheta_k}, (\mathbf{P}_j^{\vartheta_k})^{-1}\}$$

matching the expectation $\boldsymbol{\mu}_j^{\vartheta_k}$ and precision matrix $\mathbf{P}_j^{\vartheta_k}$ with the mode and the inverse curvature at the mode of the full conditional. Note that the mode and curvature are evaluated for the logarithm of the full conditional density such that the normalizing constant is not required. This yields

$$\boldsymbol{\mu}_j^{\vartheta_k} = (\mathbf{P}_j^{\vartheta_k})^{-1}(\mathbf{Z}_j^{\vartheta_k})'\mathbf{W}^{\vartheta_k}(\mathbf{z}^{\vartheta_k} - \boldsymbol{\eta}_{-j}^{\vartheta_k}) \qquad \mathbf{P}_j^{\vartheta_k} = (\mathbf{Z}_j^{\vartheta_k})'\mathbf{W}^{\vartheta_k}\mathbf{Z}_j^{\vartheta_k} + \frac{1}{(\tau_j^{\vartheta_k})^2}\mathbf{K}_j^{\vartheta_k} \tag{3}$$

where $\mathbf{z}^{\vartheta_k} = \boldsymbol{\eta}^{\vartheta_k} + (\mathbf{W}^{\vartheta_k})^{-1}\mathbf{v}^{\vartheta_k}$ are the working observations, $\boldsymbol{\eta}_{-j}^{\vartheta_k}$ denotes the predictor without the jth component, $\mathbf{v}^{\vartheta_k} = \partial l/\partial \boldsymbol{\eta}^{\vartheta_k}$ are the score vectors of the log-likelihood $l \equiv l(\boldsymbol{\eta}^{\vartheta_1}, \ldots, \boldsymbol{\eta}^{\vartheta_K})$ and $\mathbf{W}^{\vartheta_k}$ are working weight matrices, $w_i^{\vartheta_k}$ equals $\mathbb{E}\{-\partial^2 l/(\partial \eta_i^{\vartheta_k})^2\}$ on the diagonals and 0 otherwise. As a consequence, the generic MCMC algorithm is applicable whenever the first and second derivative of the log-likelihood exist (which coincides with one of the standard assumptions for maximum likelihood theory). The main issue when considering a new distribution is to derive these derivatives (and potentially the expectations of the second derivative). If this is not possible analytically, appropriate approximations can be used to construct the proposal density. Note that the full conditionals of the smoothing variances $p\{(\tau_j^{\vartheta_k})^2|\cdot\}$ are again inverse gamma distributions with parameters

$$\begin{aligned} \tilde{a}_j^{\vartheta_k} &= \tfrac{1}{2}rk(\mathbf{K}_j^{\vartheta_k}) + a_j^{\vartheta_k}, \\ \tilde{b}_j^{\vartheta_k} &= \tfrac{1}{2}(\boldsymbol{\beta}_j^{\vartheta_k})'\mathbf{K}_j^{\vartheta_k}\boldsymbol{\beta}_j^{\vartheta_k} + b_j^{\vartheta_k}. \end{aligned} \tag{4}$$

The modularity of the MCMC method allows us to represent the sampler in a unified framework where in each iteration $t=1,\ldots,T$ the final Metropolis–Hastings algorithm loops over the regression coefficients of different effects and over all distribution parameters. In summary, the procedure proposed can therefore be seen as a multivariate extension of that given in Klein *et al*. (2013) for the univariate case.

Step 1: for $t=1,\ldots,T$ go to step 2.
Step 2: for $k=1,\ldots,K$ go to step 3.
Step 3: for $j=1,\ldots,J_k$ propose $\boldsymbol{\beta}_j^p$ from the density $q\{(\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}, \boldsymbol{\beta}_j^p\} = N[(\boldsymbol{\mu}_j^{\vartheta_k})^{[t]}, \{(\mathbf{P}_j^{\vartheta_k})^{[t]}\}^{-1}]$ with expectation $\boldsymbol{\mu}_j^{\vartheta_k}$ and precision matrix $\mathbf{P}_j^{\vartheta_k}$ given in expression (3) and accept $\boldsymbol{\beta}_j^p$ as a new state of $(\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}$ with acceptance probability

$$\alpha\{(\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}, \boldsymbol{\beta}_j^p\} = \min\left[\frac{p(\boldsymbol{\beta}_j^p|\cdot)\, q\{\boldsymbol{\beta}_j^p, (\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}\}}{p\{(\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}|\cdot\}\, q\{(\boldsymbol{\beta}_j^{\vartheta_k})^{[t]}, \boldsymbol{\beta}_j^p\}}, 1\right].$$

If $\mathbf{K}_j^{\vartheta_k} \neq \mathbf{0}$ draw a new state for $(\tau_j^{\vartheta_k})^2$ from an inverse gamma distribution with parameters given in expression (4).

In both our applications and the simulations that are documented in Section C of the online supplement, we observed satisfactory mixing and convergence properties of our generic algorithm, i.e. high acceptance rates roughly between 70% and 100% as well as small autocorrelations. As an additional investigation of the properties of our MCMC approach, we conducted a simulation on the behaviour under concurvity, i.e. situations in which the regression functionals are correlated with each other; see section D of the supplement for detailed documentation. The basic outcome here is that the behaviour is fairly robust against concurvity and strong dependences of the MCMC chains for different effects are only observed in the most extreme scenario when one functional effect is a perfect deterministic linear combination of the other effects of the predictor.

2.4. Model choice

In practical applications of multivariate distributional regression, we are typically faced with the necessity of several model selection decisions concerning the choice of a response distribution as well as the choice of relevant influential variables for the different predictors in a model. Owing to the high flexibility of multivariate distributional regression, any model choice approach will be rather time consuming and we therefore discuss some guidelines that facilitate model choice even for large data sets and many explanatory variables as in the childhood undernutrition example. For the future, automatic model choice approaches like that of Scheipl *et al.* (2012) utilizing spike and slab prior structures would be quite attractive but are beyond the scope of this paper.

For model choice in multivariate distributional regression, we rely on the deviance information criterion (DIC) (Spiegelhalter *et al.*, 2002), normalized quantile residuals (Dunn and Smyth, 1996) and proper scoring rules (Gneiting and Raftery, 2007). In principle, all three can be used for discriminating between competing response distributions and predictor structures of the distribution parameters. Still, we consider each of the criteria to be most useful for specific aspects of the model choice procedure: Whereas the DIC is based on the deviance and the effective number of parameters and therefore provides a compromise between fidelity to the data and model complexity, quantile residuals are more a graphical device to check the general fit of the data under a specific type of response distribution, and proper scoring rules are summary measures for the evaluation of probabilistic forecasts. Furthermore, although computation of the DIC is part of the estimation run, the scoring rules require predictive distributions derived from the estimated models.

We now discuss each of the criteria for model choice in more detail.

2.4.1. Quantile residuals

The concept of quantile residuals that was used by Klein *et al.* (2013) in the context of univariate Bayesian structured additive distributional regression is unfortunately not directly extendable to the multivariate framework since the multivariate cumulative distribution function F cannot be inverted, i.e. $F(\mathbf{y}|\boldsymbol{\vartheta})$, $\boldsymbol{\vartheta} = (\vartheta_1, \ldots, \vartheta_K)$, is no longer uniformly distributed. However, quantile residuals can still be applied to the marginal distributions to evaluate at least the marginal fit of

the model. Note that an adequate marginal model fit is a necessary but not a sufficient condition for a satisfactory fit of the multivariate model.

To determine marginal quantile residuals, we use the estimated distribution parameters that are summarized in $\hat{\boldsymbol{\vartheta}}$ to compute the quantile residuals $\hat{r}_d = \Phi^{-1}\{\hat{F}_d(y_d|\hat{\boldsymbol{\vartheta}})\}$ of the marginal distributions $\hat{F}_d$, $d = 1, \ldots, D$, and where Φ^{-1} is the inverse distribution function of a standard normal distribution. If $\hat{F}_d$ is close to the true marginal distribution, the quantile residuals are approximately standard normally distributed and can therefore be visualized in quantile–quantile plots. In our experience, quantile residuals are quite robust concerning the exact specification of different predictors for a fixed response distribution such that the choice of the response distribution can, for example, be based on a reasonably complex model specification. We shall come back to this issue in Section 3.

2.4.2. Deviance information criterion

To obtain a best fitting model given a response distribution, we perform a stepwise forward selection based on the DIC. To avoid a full stepwise procedure over all distributional parameters simultaneously, we define an order for the parameters and optimize them in this order. For example, in the case of the normal distribution, we start with the expectations, consider the standard deviations next and finally optimize the correlation parameter. Still, the flexibility in model specifications may tempt researchers to specify overly complex models. To aim at favouring sparse models, small differences in DIC values of two competing models can be assisted by looking at significances of the estimated effects. In such situations we propose to exclude parametric effects if the 95% credible interval contains zero. For non-parametric effects we choose the 95% simultaneous credible band constructed following Krivobokova *et al*. (2010).

2.4.3. Proper scoring rules

The predictive ability of a chosen model compared with a competing model can be evaluated with proper scoring rules. A scoring rule assigns a score $S(F, \mathbf{y})$ to each pair $(F, \mathbf{y})$ where F is the predictive distribution and $\mathbf{y} = (y_1, \ldots, y_D)' \in \mathbb{R}^D$ is the observation vector. Following Gneiting and Raftery (2007), we define a scoring rule S to be proper if $\mathbb{E}_{F_0}\{S(F, \mathbf{y})\} \leqslant \mathbb{E}_{F_0}\{S(F_0, \mathbf{y})\}$ for all F_0 and F and it is strictly proper if equality only holds if $F = F_0$. In practice, we obtain the scores based on k-fold cross-validation. Specific scoring rules will be discussed in Section 3.

3. Chronic and acute forms of childhood undernutrition in India

Childhood undernutrition is one of the most urgent public health challenges in developing and transition countries since it is not only related to the growth of children but also has severe long-term socio-economic and health effects (UNICEF, 1998). A rich database with information on fertility, family planning and maternal and child health, as well as child survival, human immunodeficiency virus and acquired immune deficiency syndrome, malaria and nutrition is provided by demographic and health surveys (`www.measuredhs.com`) consisting of more than 300 surveys conducted in 90 countries. Usually, childhood undernutrition is measured by a Z-score that compares the nutritional status of children in the population of interest with the nutritional status in a reference population. Consequently the Z-score is defined as

$$Z_i = \frac{\mathrm{AC}_i - \mu_{\mathrm{AC}}}{\sigma_{\mathrm{AC}}}$$

where AC_i denotes the anthropometric characteristic for child i, whereas μ_{AC} and σ_{AC} correspond to the median and standard deviation in the reference population (stratified with respect

to age and gender). Depending on the choice of the anthropometric indicator, different aspects of undernutrition can be assessed. Insufficient height for age is an indicator for chronic undernutrition (*stunting*) whereas insufficient weight for height captures acute undernutrition (*wasting*). Here, we focus on the joint model with a special interest in the correlation between both scores which is of considerable relevance for policy makers who are concerned about undernutrition. In particular, a positive correlation would suggest particular urgency to address undernutrition among children who suffer from both acute and chronic forms of undernutrition. It would also imply that addressing one of the issues might help in addressing the other. Counteracting this is a negative correlation that is expected from a biological perspective, particularly at younger ages. If a child is born with adequate height, reduced nutrient intake will lead immediately to wasting, thus generating a negative correlation. It will also stop growing, thus slowly becoming stunted, which will make it somewhat easier to maintain adequate weight for its (reduced) height, implying a negative correlation between changes in the two indicators over time. Over time, however, continued nutritional problems related to low caloric intake and the presence of disease might weaken this negative correlation by negatively affecting growth and weight gain at the same time.

Our analysis is based on the 1998–1999 survey for India, containing information on 24 316 children (after deleting implausible and incomplete observations). India is a particularly interesting case since it is among the countries with the highest rates of childhood undernutrition in the world, with high rates of both stunting and wasting (Klasen and Moradi, 2000; Klasen, 2008). A detailed description and a preselection of the large number of all potential covariates that are provided in the data set is given in Belitz and Lang (2008), Fahrmeir and Kneib (2011) or Belitz *et al*. (2010). A boosting and quantile regression based analysis of a similar data set from the 2006 India demographic and health survey without spatial information can be found in Fenske *et al*. (2011).

3.1. Model choice

As possible response distributions, we consider the bivariate normal and the t-distribution and follow the steps that were proposed in Section 2.4 to select optimal models.

3.1.1. Quantile residuals

We start with full models, i.e. we assume that

$$\begin{aligned}\eta_i = \mathbf{x}_i'\boldsymbol{\beta} + f_1(\text{cage}_i) + f_2(\text{breastfeeding}_i) + f_3(\text{mage}_i) + f_4(\text{mbmi}_i)\\ + f_5(\text{medu}_i) + f_6(\text{edupartner}_i) + f_{\text{spat}}(\text{dist}_i) + \beta_{\text{dist}_i}\end{aligned}$$

for all model parameters. Here, f_1–f_6 are smooth functions of the covariates age of the child (cage) in months, lactation (breastfeeding) in months, current age of the mother (mage), body mass index of the mother (mbmi), education years of the mother (medu) and education years of the mother's partner (edupartner). We apply cubic Bayesian P-splines with 20 inner knots and a second-order random-walk prior for the non-linear smooth terms of continuous covariates. The vector $\mathbf{x}_i$ contains a constant comprising the overall level of the predictors and several linear effects (e.g. binary and categorical variables; see the on-line supplement Table E1 for a complete list). The spatial function f_{spat} and the district-specific random effect β_{dist} represent the complete spatial effect of the district that the child belongs to. Whereas f_{spat} captures spatial correlations, β_{dist} catches local and small-scale variations. The former is based on an intrinsic Gaussian Markov random-field prior where two regions are treated as neighbours if they share a common boundary. The latter is assigned an independent and identically distributed Gaussian

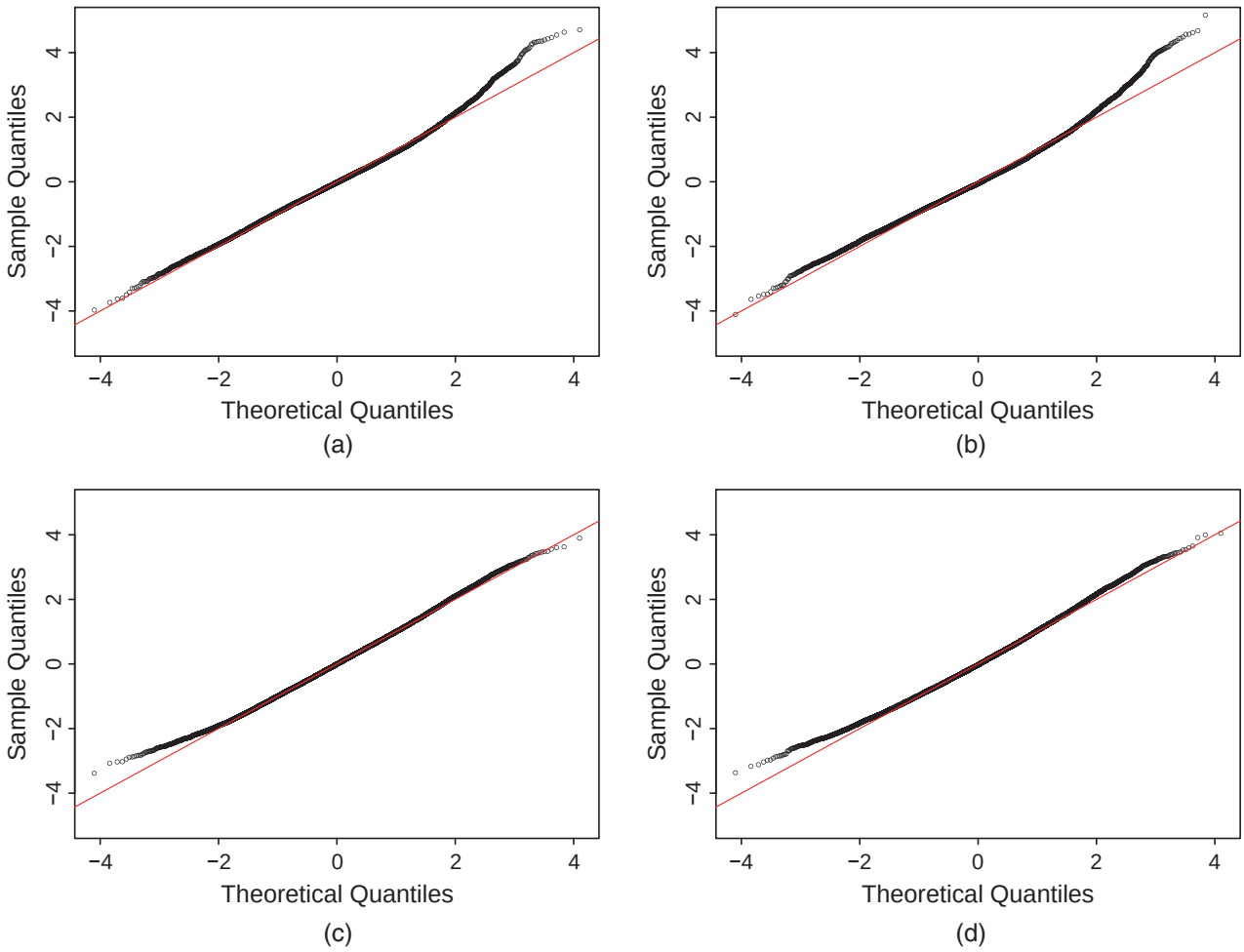

Fig. 1. Childhood undernutrition—quantile residuals of the marginal distributions in (a), (b) the full bivariate model and (c), (d) the full bivariate t-model: (a), (c) stunting; (b), (d) wasting

prior, therefore corresponding to a random intercept based on clusters defined by regions. Lang and Fahrmeir (2001) have shown through simulations that the two components of the complete spatial effect cannot in general be separated and only the sum of both effects can be estimated satisfactorily. Therefore results for spatial effects will always contain the sum of the two parts.

The resulting residuals of the marginal distributions are visualized in Fig. 1 and indicate that both distributions deliver a reasonable fit for residuals between -2 and 2. Large residuals are better captured by the t-distributions whereas the small residuals of the normal distributions are closer to the diagonal compared with the t-distributions. As mentioned before, the residuals rarely change when comparing different predictor specifications (over a reasonable set of predictors) for a fixed response distribution; see Fig. E19 of the on-line supplement where the residuals of the marginal distributions in the best fitting models are shown.

3.1.2. Deviance information criterion

We face 11 binary and categorical covariates, six continuous covariates and the spatial effect. As an order for the stepwise model selection procedure, we choose μ_1, μ_2, σ_1, σ_2 and ρ (and n_{df} in the case of the t-distribution) such that we start with the location parameters, go on with the scale parameters and end with the correlation parameter (or with the degrees of freedom). Owing to the large number of potential covariates we utilize the findings of Fahrmeir and Kneib

(2011) in the univariate regression models for the expectations μ_1 and μ_2 of stunting and wasting and proceed as follows.

(a) As in Fahrmeir and Kneib (2011) we use all available covariates to estimate the location parameters.
(b) Parametric effects are added also to the remaining predictors without being part of the selection procedure.
(c) From the full model, we find that the 95% simultaneous credible band of medupartner on σ_1 fully covers the zero line. For σ_2 this is also true for the effects mage and medu.

Since our focus is on the correlation parameter, the potential candidates for the scale parameters are the reduced number of significant effects in the full model. Significant here means that the 95% simultaneous credible band does not fully cover the zero line.

The predictors for scale and correlation parameters resulting from the stepwise selection are then

$$\left.\begin{aligned}
\eta_i^{\sigma_1} &= (\mathbf{x}_i^{\sigma_1})'\boldsymbol{\beta}^{\sigma_1} + f_1^{\sigma_1}(\text{cage}_i) + f_2^{\sigma_1}(\text{breastfeeding}_i) + f_3^{\sigma_1}(\text{mage}_i) + f_4^{\sigma_1}(\text{mbmi}_i) \\
&\quad + f_5^{\sigma_1}(\text{medu}_i) + f_{\text{spat}}^{\sigma_1}(\text{dist}_i) + \beta_{\text{dist}_i}^{\sigma_1}, \\
\eta_i^{\sigma_2} &= (\mathbf{x}_i^{\sigma_2})'\boldsymbol{\beta}^{\sigma_2} + f_1^{\sigma_2}(\text{cage}_i) + f_4^{\sigma_2}(\text{mbmi}_i) + f_{\text{spat}}^{\sigma_2}(\text{dist}_i) + \beta_{\text{dist}_i}^{\sigma_2}, \\
\eta_i^{\rho} &= (\mathbf{x}_i^{\rho})'\boldsymbol{\beta}^{\rho} + f_1^{\rho}(\text{cage}_i) + f_{\text{spat}}^{\rho}(\text{dist}_i) + \beta_{\text{dist}_i}^{\rho}, \\
\eta_i^{n_{\text{df}}} &= (\mathbf{x}_i^{n_{\text{df}}})'\boldsymbol{\beta}^{n_{\text{df}}} + f_{\text{spat}}^{n_{\text{df}}}(\text{dist}_i) + \beta_{\text{dist}_i}^{n_{\text{df}}}
\end{aligned}\right\} \tag{5}$$

for the bivariate t-distribution and

$$\begin{aligned}
\eta_i^{\sigma_1} &= (\mathbf{x}_i^{\sigma_1})'\boldsymbol{\beta}^{\sigma_1} + f_1^{\sigma_1}(\text{cage}_i) + f_2^{\sigma_1}(\text{breastfeeding}_i) + f_3^{\sigma_1}(\text{mage}_i) + f_4^{\sigma_1}(\text{mbmi}_i) \\
&\quad + f_5^{\sigma_1}(\text{medu}_i) + f_{\text{spat}}^{\sigma_1}(\text{dist}_i) + \beta_{\text{dist}_i}^{\sigma_1}, \\
\eta_i^{\sigma_2} &= (\mathbf{x}_i^{\sigma_2})'\boldsymbol{\beta}^{\sigma_2} + f_1^{\sigma_2}(\text{cage}_i) + f_4^{\sigma_2}(\text{mbmi}_i) + f_{\text{spat}}^{\sigma_2}(\text{dist}_i) + \beta_{\text{dist}_i}^{\sigma_2}, \\
\eta_i^{\rho} &= (\mathbf{x}_i^{\rho})'\boldsymbol{\beta}^{\rho} + f_1^{\rho}(\text{cage}_i) + f_2^{\rho}(\text{breastfeeding}_i) + f_3^{\rho}(\text{mage}_i) + f_{\text{spat}}^{\rho}(\text{dist}_i) + \beta_{\text{dist}_i}^{\rho}
\end{aligned}$$

for the bivariate normal distribution. The fact that only some of the mothers' characteristics remain in the model after model selection is related to the fact that many of them are correlated so that the addition of more characteristics does not add to the explanatory power of the regression. See Caputo *et al.* (2003) for an analysis using graphical chain models to uncover the many direct and indirect pathways of the effect of mother's characteristics on childhood undernutrition. The DIC differences between the full models and the models that were selected with the DIC are quite small for both distributions. Absolute values of the DIC indicate a slight preference for the reduced bivariate t-model (5); see Table 1.

3.1.3 Proper scoring rules.

We consider three proper scoring rules based on the quadratic score $S(p_r, \mathbf{y}_r) = 2p_r(y_{r1}, y_{r2}) - \|p_r(y_{r1}, y_{r2})\|_2^2$, the spherical score $S(p_r, \mathbf{y}_r) = p_r(y_{r1}, y_{r2})/\|p_r(y_{r1}, y_{r2})\|_2$ and the logarithmic score $S(p_r, \mathbf{y}_r) = \log\{p_r(y_{r1}, y_{r2})\}$ under the Lebesgue measure on the measurable space $(\Omega, \mathcal{F}) = (\mathbb{R}^2, \mathcal{B}(\mathbb{R}^2))$ and with

$$\|p_r(x, y)\|_2 = \left\{ \int \int |p_r(x_1, x_2)|^2 \, \mathrm{d}x_1 \, \mathrm{d}x_2 \right\}^{1/2}.$$

Here, $\mathbf{y}_r = (y_{r1}, y_{r2})'$ is an observation from a hold-out sample $\mathbf{y}_1, \ldots, \mathbf{y}_R$ and p_r is the density of a bivariate normal or a bivariate t-distribution with plugged-in parameters obtained from

Table 1. Childhood undernutrition: comparison of the DIC achieved from estimates based on the whole data set and average score contributions obtained from tenfold cross-validations†

Distribution	*DIC*	*Quadratic score*	*Logarithmic score*	*Spherical score*
Normal (selected model)	157690	0.06346	−3.23833	0.25061
Normal (full model)	157685	0.06344	−3.23906	0.25058
t (selected model)	*156446*	*0.06384*	*−3.22208*	*0.25111*
t (full model)	156462	*0.06384*	−3.22286	*0.25111*

†Full models are the models with all covariates estimated in all parameters. The models selected (in italics) correspond to the models resulting from the forward selections.

the estimation data of the current cross-validation fold. More specifically, we split the data set into 10 equal parts, use nine parts for estimation and predict the parameters for the remaining part. The predictive ability of the two models can then be compared by the aggregated average score $S_R = (1/R)\Sigma_{r=1}^{R} S(F_r, \mathbf{y}_r)$ with predictive distributions

$$F_r(y_{r1}, y_{r2}) = \int_{-\infty}^{y_{r1}} \int_{-\infty}^{y_{r2}} p_r(x_1, x_2)\, \mathrm{d}x_1 \mathrm{d}x_2.$$

The average scores over the 10 folds for the full models and selected models under the bivariate normal and the bivariate *t*-distribution are given in Table 1. The propriety of the scores ensures that higher scores deliver better forecasts when comparing two competing models. Similarly to the residuals and DIC values, the differences in averaged scores are quite small but marginally in favour of the *t*-distribution.

In summary, all goodness-of-fit measures slightly favour the model resulting from our forward selection procedure in the bivariate *t*-model. In addition, the estimated effects based on the bivariate normal and *t*-distribution are visually close to each other. Therefore, we restrict our presentation primarily to the results that are based on the bivariate *t*-distribution.

For all models, we used 52 000 MCMC iterations with a burn-in phase of 2000 iterations and a thinning parameter of 50 and observed high acceptance rates and small auto-correlations; see Fig. E20 of the on-line supplement for two exemplary sampling paths of coefficients in the *t*-model selected.

3.2. Results

A complete presentation and discussion of results would go beyond the scope of the paper but additional results can be found in the on-line supplement section E where also the results of the categorical variables are presented. Here, we primarily focus on the results for the correlation parameter between stunting and wasting.

Fig. 2 visualizes posterior mean point estimates and 80% posterior probabilities of the complete spatial effect f_{spat}^{ρ} on the correlation parameter ρ. Fig. 3 shows posterior mean estimates of the selected non-linear effect $f_1^{\rho}(\text{cage})$ on the predictor level together with 80% and 95% pointwise credible intervals (Fig. 3(a)) and the estimated correlation as a function of cage (Fig. 3(b)), both adjusted for the overall constant β_0^{ρ}. The estimated effects of binary and categorical covariates can be found in Table E6 of the on-line supplement. In Figs 4 (selected normal model) and 5 (selected *t*-model) contour lines of estimated densities for four different ages are depicted.

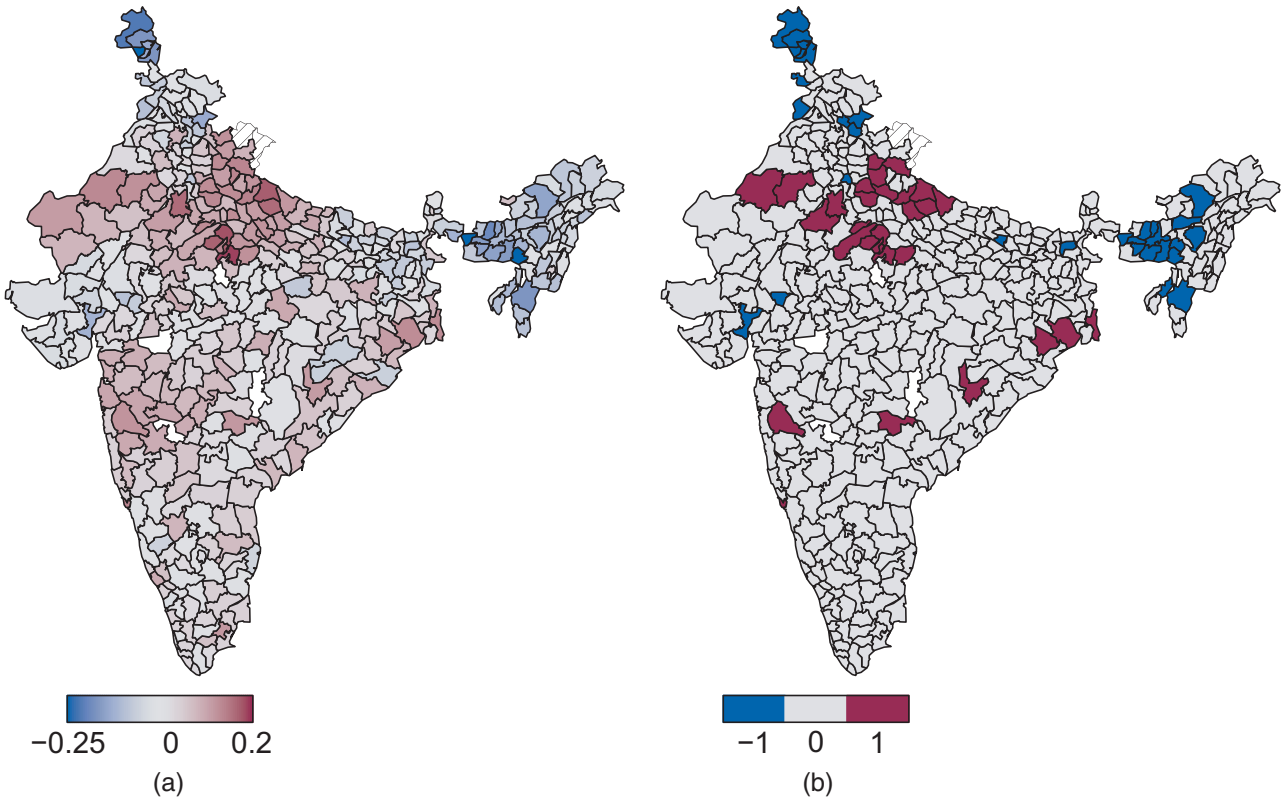

Fig. 2. Childhood undernutrition, selected bivariate t-model—posterior mean estimates of the complete spatial effects on (a) ρ and (b) 80% posterior probabilities: in (b) a value of 1 corresponds to a strictly positive 80% credible interval and a value of -1 to a strictly negative credible interval; a value of 0 indicates that the corresponding credible interval contains 0; note that in (a) the plot range is not centred at 0 for visibility

The densities are obtained by plugging in posterior mean estimates of the distribution parameters as follows: the remaining non-linear effects are evaluated at average covariate values, i.e. at $f(\bar{x})$. Binary and categorical variables are set to 0; the reference levels can be found in Table E1 of the on-line supplement.

Especially in the north of India, particularly in districts at the border with China and Bhutan, several districts have a negative effect on the correlation whereas in some districts in the centre of the country we estimate a small positive effect; see Fig. 2. These effects could suggest that in these central parts there is an acute undernutrition problem in the year of the survey, in addition to an existing chronic problem, requiring particular intervention to address this combination of problems. Fig. 3 shows that the effect of age of the child on the correlation between stunting and wasting turns out to be non-linear but negative throughout. But note the steady increase until an age of about 2 years so the negative correlation is much weaker then.

The contour plots in Figs 4 and 5 also indicate that the absolute value of the correlation between wasting and stunting decreases with the age of the child; it is highly negative at 3 and 6 months of age, and much weaker (but still negative) at 2 years. Note also that the mode of the contour plot moves with age, particularly for stunting. Whereas at age 3 months the mode is around a Z-score of about -0.5, it moves to a Z-score of about -2 by an age of 2 years. Furthermore, the plots show that the variance of stunting increases whereas that of wasting becomes smaller for older children. Related to the discussion above, this suggests that, at young ages, a negative correlation between stunting and wasting appears to dominate. In particular, it appears that, at birth, those children with greater height suffer from lower weight for height and vice versa. This could be linked to the fact that there are upper biological limits to birth weight

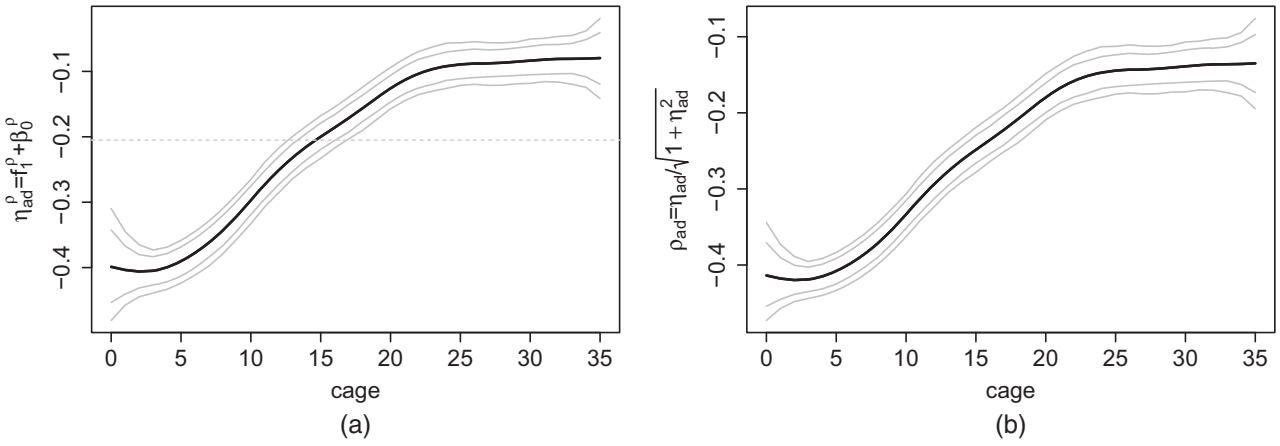

Fig. 3. Childhood undernutrition, selected bivariate *t*-model: (a) posterior mean estimates of the selected non-linear effect f^{ρ}_1(cage) on the predictor level adjusted by the overall constant β^{ρ}_0 (i.e. $\eta^{\rho}_{ad} = f^{\rho}_1(\text{cage}) + \beta^{\rho}_0$) and together with 80% and 95% pointwise credible intervals (the effect f^{ρ}_1 is centred at 0 ($\Sigma_i f^{\rho}_i(\text{cage}) = 0$); (b) posterior mean estimates of the adjusted correlation parameter ρ_{ad}(stunting, wasting) = $\eta^{\rho}_{ad}/\sqrt{\{1 + (\eta^{\rho}_{ad})^2\}}$ as a function of the selected non-linear effect cage together with 80% and 90% pointwise credible intervals (the other effects are set to 0)

that can be sustained, particularly when mothers also suffer from undernutrition and have small body stature, as is prevalent in India. Thus mothers who turn out to have (genetically) taller children deliver them with a lower weight for height than mothers with shorter children. This negative correlation then slowly dissipates as tall children then fail to grow and become stunted. They then also become more susceptible to socio-economic and environmental conditions that affect both weight and height, thus counteracting the negative link that was prevalent at birth or is related to the dynamic interaction between the two variables (see Wiesenfarth *et al.* (2012) for further discussion). The densities that are obtained from a *t*-distribution are more flat and with heavier tails as expected.

Table E6 of the on-line supplement shows that socio-economic effects are associated with an increase in the correlation between the stunting and wasting *Z*-score. This is interesting and supports the argument that was made above that socio-economic factors generate a positive correlation between stunting and wasting; as children age, these socio-economic factors appear to become more important, reducing the initial negative correlation between the two indicators.

In a nutshell, the estimates indicate that there is an important dynamic interaction between stunting and wasting that changes as the child ages. Whereas stunting becomes a more permanent condition for many children, wasting stays more stable. Clearly it is valuable to study the correlation between both indicators, which provides additional insights for research and policy. In particular, the correlations suggest that at young ages it might be particularly the tall and thin babies that are of concern, whereas, among older children, the stunted children might require particular support.

4. Selected sociodemographic factors on Germany's federal elections

In this section, we present an analysis on Germany's federal election in 2009 as an example of Dirichlet regression. The data were provided by Statistische Ämter des Bundes und der Länder (`www.destatis.de`) and contain proportions of the electorate voting (the response variable) on five parties for each of the 413 districts (*Landkreise*) in Germany. The proportion of votes for the Christian Democratic Union and Christian Social Union, Social Democratic Party, the Liberals, the Left, the Greens and others sum to 1 in each district. As covariates, we consider

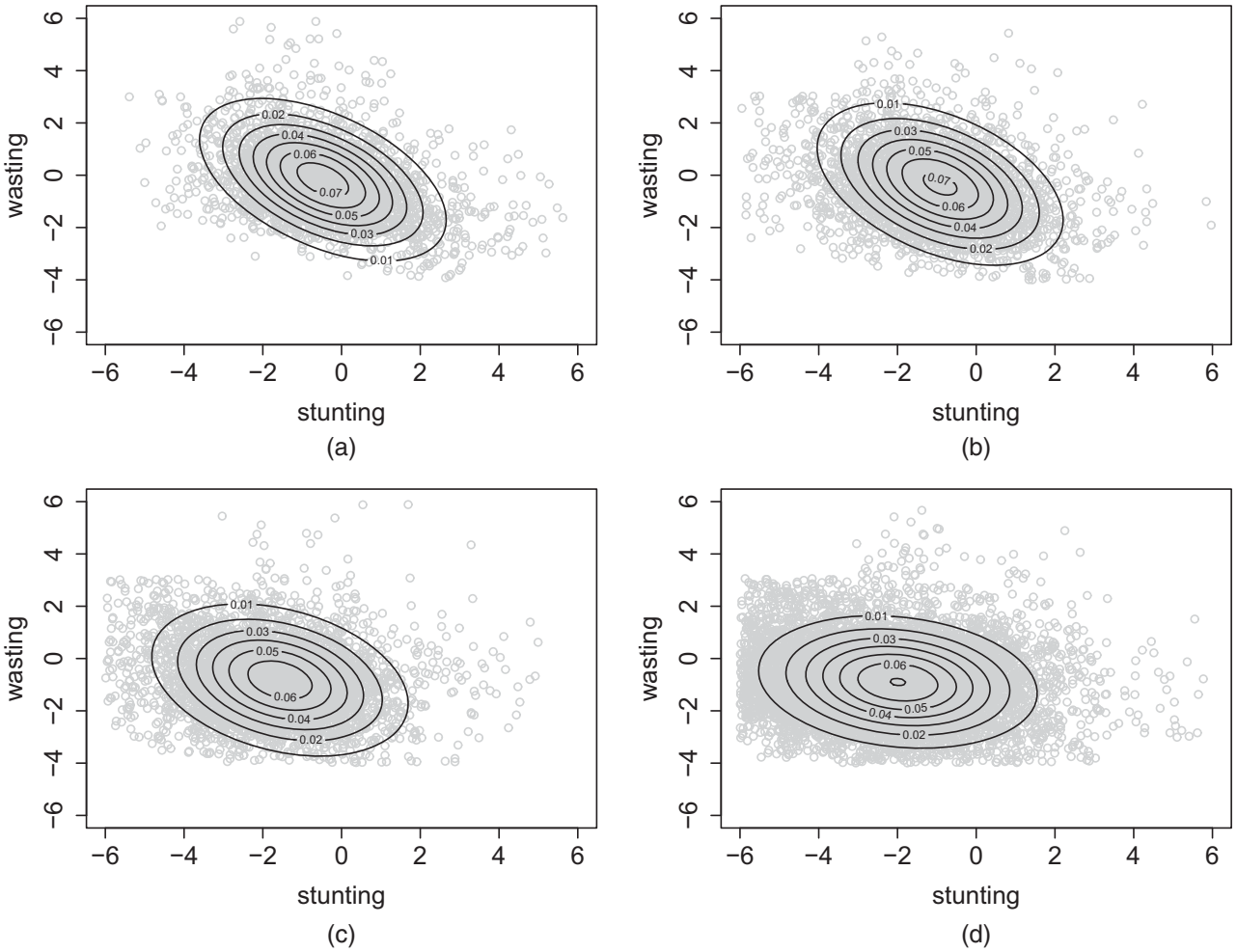

Fig. 4. Childhood malnutrition, selected bivariate normal model contour lines of densities for four different ages (a) 3 months ($\rho = -0.49$; $\mu_{\text{stunting}} = -0.48$; $\mu_{\text{wasting}} = -0.17$; $\sigma_{\text{stunting}} = 1.55$; $\sigma_{\text{wasting}} = 1.54$; ○, children with ages between 0 and 4.5 months), (b) 6 months ($\rho = -0.44$; $\mu_{\text{stunting}} = -0.92$; $\mu_{\text{wasting}} = -0.34$; $\sigma_{\text{stunting}} = 1.57$; $\sigma_{\text{wasting}} = 1.56$; ○, children with ages between 4.5 and 9 months), (c) 9 months ($\rho = -0.31$; $\mu_{\text{stunting}} = -1.56$; $\mu_{\text{wasting}} = -0.82$; $\sigma_{\text{stunting}} = 1.66$; $\sigma_{\text{wasting}} = 1.49$; ○, children with ages between 9 and 15 months) and (d) 24 months ($\rho = -0.17$; $\mu_{\text{stunting}} = -1.99$; $\mu_{\text{wasting}} = -0.9$; $\sigma_{\text{stunting}} = 1.79$; $\sigma_{\text{wasting}} = 1.28$; ○, children with ages between 15 and 36 months): the remaining non-linear effects are kept constant at $f_j(\bar{x}_j)$ (estimated functions evaluated at mean covariate values); the binary and categorical variables are set to 0

district-specific quantities, i.e. the proportion of electorates (PoE) compared with the population entitled to vote (the turnout) in per cent, the rate of unemployment (unemployment) in 2008, the gross domestic product *per capita* (GDPpc) in 2008 (measured in thousand euros) and one of 38 administrative regions (region) that the districts are in. The predictors $\eta_i^{\alpha_d} = \log(\alpha_{id})$ for $d = 1, \ldots, 6$ and $i = 1, \ldots, 413$ are hence of the form

$$\eta_i = \beta_0 + f_1(\text{PoE}_i) + f_2(\text{GDPpc}_i) + f_3(\text{unemployment}_i) + f_{\text{spat}}(\text{region}_i). \tag{6}$$

As described in Section 3, f_1–f_3 are smooth functions modelled with cubic Bayesian P-splines with 20 inner knots and second-order random-walk prior and f_{spat} is assigned a Markov random-field prior on 38 administrative regions in Germany.

Since each of the covariates is important for predicting at least one of the parties, we did not perform model choice in this example. Selecting optimal predictors separately for the parameters α_d would lead to models that contain different explanatory variables for the different parties. This would make interpretation more challenging. As a consequence, we shall always use predictor

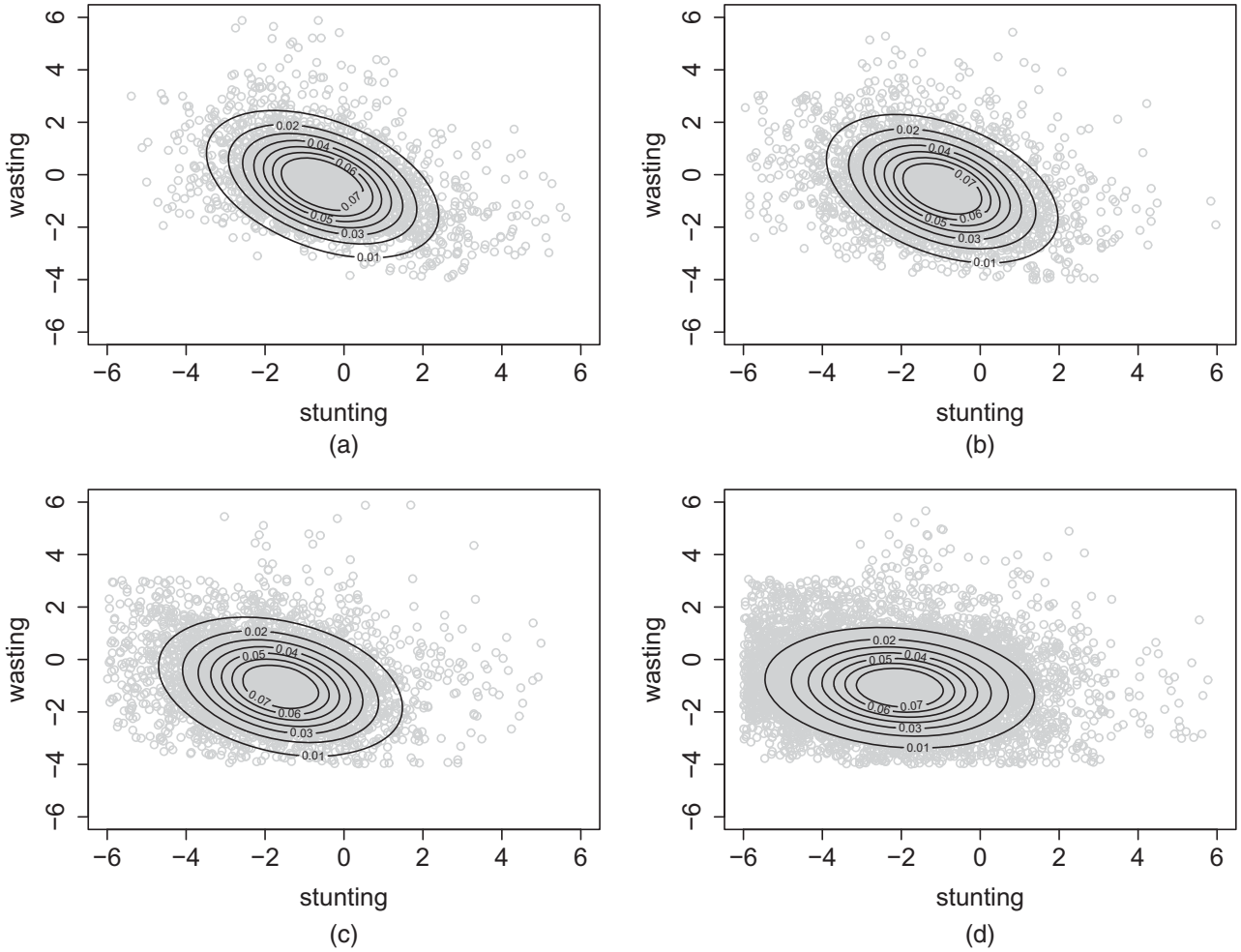

Fig. 5. Childhood nutrition, selected bivariate t-model ($n_{df} = 10.48$)—contour lines of densities for four different ages (a) 3 months ($\rho = -0.42$; $\mu_{stunting} = -0.54$; $\mu_{wasting} = -0.36$; $\sigma_{stunting} = 1.37$; $\sigma_{wasting} = 1.3$; ○, children with ages between 0 and 4.5 months), (b) 6 months ($\rho = -0.4$; $\mu_{stunting} = -0.96$; $\mu_{wasting} = -0.54$; $\sigma_{stunting} = 1.37$; $\sigma_{wasting} = 1.33$; ○, children with ages between 4.5 and 9 months), (c) 9 months ($\rho = -0.29$; $\mu_{stunting} = -1.61$; $\mu_{wasting} = -1.04$; $\sigma_{stunting} = 1.47$; $\sigma_{wasting} = 1.26$; ○, children with ages between 9 and 15 months) and (d) 24 months ($\rho = -0.15$; $\mu_{stunting} = -2.04$; $\mu_{wasting} = -1.08$; $\sigma_{stunting} = 1.62$; $\sigma_{wasting} = 1.09$; ○, children with ages between 15 and 36 months): the remaining non-linear functions are kept constant at $f_j(\bar{x}_j)$ (estimated effects evaluated at mean covariate values); the binary and categorical variables are set to 0

(6) in what follows. To assess the fit of the model, we also computed marginal quantile residuals for the marginal beta distributions that are implied by the model (see Fig. F30 in the on-line supplement). All marginal residuals approximately follow straight lines but deviate more or less strongly from the diagonal line. As a consequence, our model seems to fit the general form of the distribution well but fails to model the variability precisely.

To ensure convergence, we used 102 000 MCMC iterations with a burn-in phase of 2000 iterations and a thinning parameter of 100. The acceptance rates for all effects are higher than 80% and the auto-correlations are small; see Fig. F31 of the on-line supplement for two exemplary sampling paths of coefficients.

For better interpretation, we computed expected proportions $\exp(\alpha_k)/\Sigma_{d=1}^{6}\exp(\alpha_d)$ of votes for each party and for every effect whereas all other effects are constant with estimated effects evaluated at mean covariate values. In Fig. 6, the expected values can be compared for various regions in Germany and, in Fig. 7, the expected proportions are shown in dependence of the covariates PoE, GDPpc and unemployment.

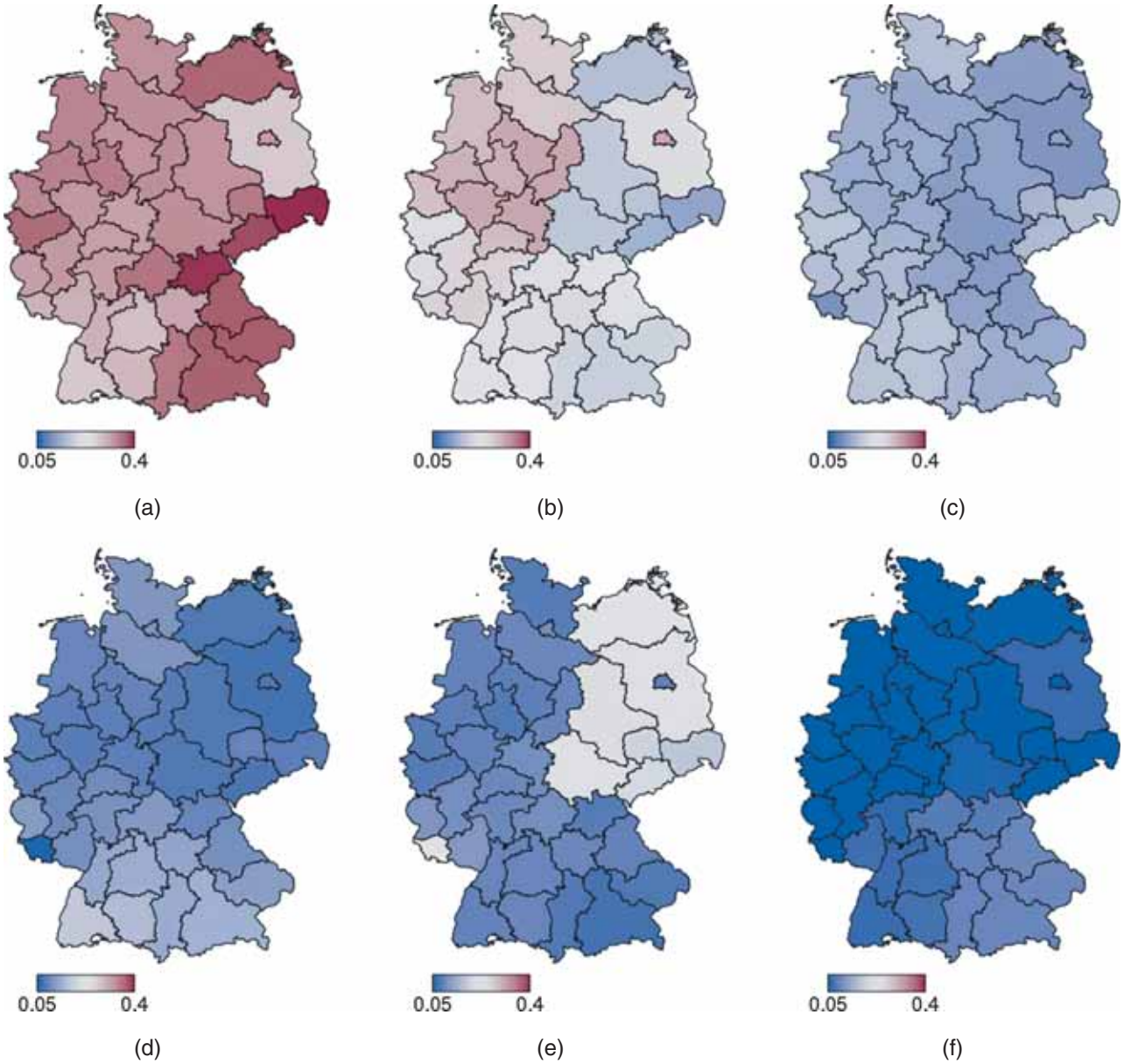

Fig. 6. Federal election—posterior mean proportions of votes in each region (all other effects are evaluated at mean covariate effects; the geometric information was provided by the Bundesinstitut für Bau-, Stadt- und Raumforschung (2002): (a) Christian Democratic Union and Christian Social Union; (b) Social Democratic Party; (c) Liberals; (d) the Greens; (e) the Left; (f) others

For the Christian Democratic Union and Christian Social Union we estimate the highest absolute proportions of votes with particular support in the east and south-east and smallest spatial variations (i.e. smallest variances τ^2; see Table F8 of the on-line supplement). Votes for the Liberals vary relatively little by region whereas for the Social Democrat Party we observe a higher constituency in the Western part of Germany compared with the rest of Germany. The Greens seem to have most votes in the south-west (since 2011 the Prime Minister of Baden-Württemberg has been provided by the Greens) whereas the most notable spatial effect is recorded for the Left. For historic reasons the Left is traditionally very strong in the east of Germany as well as the Saarland because of the prominent position of its former chairman, Oskar Lafontaine, in this federal state.

The effect of GDPpc is most pronounced for the Greens with a steady increase up to 50 000 which can be explained by their stronger support in urban areas which complies with political

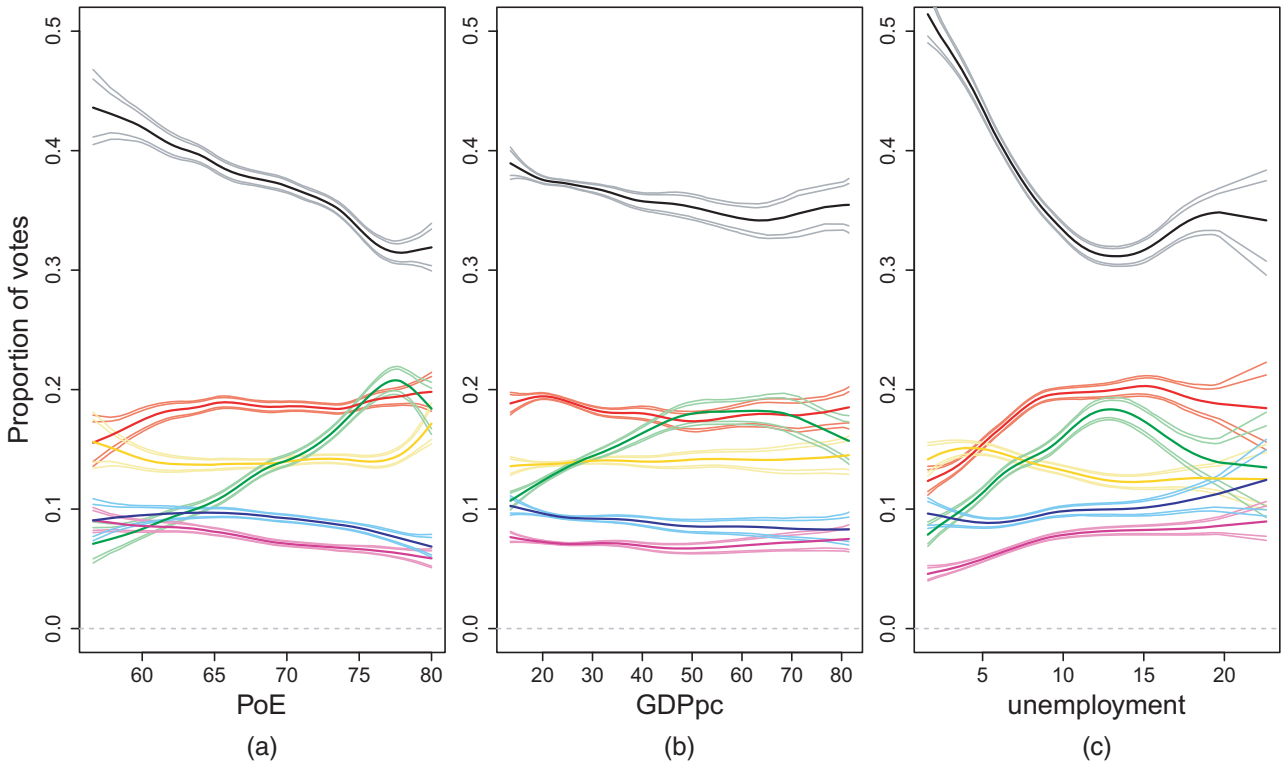

Fig. 7. Federal election—posterior mean proportions of votes together with 80% and 95% pointwise credible intervals as functions of the covariates (a) PoE (turnout in per cent), (b) GDPpc (gross domestic product *per capita* in thousand euros) and (c) unemployment (in per cent) (all others effects are evaluated at mean covariate effects): ——, Christian Democratic Union and Christian Social Union; ——, Social Democratic Party; —— Liberals; ——, Greens; ——, the Left; ——, others

literature; see for example Walter (2010). Whereas for the Social Democrat Party, Liberals and others the effect of PoE does not have clear trends, the expected proportion of votes increases for the Greens and decreases for the Christian Democratic Union and Christian Social Union as well as the Left when the voter turnout rises. Looking at unemployment, the Social Democrat Party and the Left show similar behaviour where, up to a rate of 10%, increasing unemployment is estimated to have a positive effect on votes for these two parties. For the Greens we observe the same tendency but with a negative effect for high unemployment. In contrast, the Christian Democratic Union reveals most electorates in regions with low rates of unemployment.

In conclusion it can be said that even with data on a highly aggregated level (we do not have any individual information) clear trends and differences between the parties considered can be identified and explained by the four covariates included; geographical region, turnouts, rate of unemployment and the gross domestic product *per capita*.

5. Conclusions and outlook

In this paper, we have proposed a generic Bayesian framework for structured additive distributional regression with various types of multivariate responses. The flexibility of the approach allows us to gain detailed insights into the joint stochastic behaviour of response vectors accounting for a variety of complex regression effects. We restricted our attention to pure main effects models but two-way interactions (included either as varying coefficients or based on interaction surfaces) are supported by our framework. We refrained from exploring such interactions since they would lead to even more challenging model choice problems and would also make the interpretation of estimated effects more difficult.

Although, in general, interpretation remains feasible for bivariate models for continuous responses, the truly multivariate case with dimensions higher than 2 remains difficult since complex restrictions on the parameter space must be enforced for positive definite dispersion matrices. The modified Cholesky decomposition $\boldsymbol{\Sigma} = \mathbf{L}\mathbf{D}^2\mathbf{L}'$ that was considered in Pourahmadi (2011) seems to be most promising in this regard since the non-redundant entries of $\mathbf{L}^{-1}$ are unconstrained whereas the diagonal elements in $\mathbf{D}$ correspond to the standard deviations. In the future it would thus be interesting to investigate further the findings of Pourahmadi (2011) in the context of multivariate distributional regression.

In any case, we believe that multivariate distributional regression is an important contribution to the toolbox of applied statisticians with a variety of applications. In particular, the distributional variant of seemingly unrelated regression models with all parameters depending on covariates provides a natural counterpart to recent attempts to define bivariate quantile regression models; see for example Chakraborty (2003) who considered linear regression for multivariate geometric quantiles or Chaudhuri (1996) based on a transformation–retransformation method. Albeit its admittedly much stronger assumption of a particular response distribution, the multivariate normal and the multivariate t-model have the considerable advantage of defining a coherent interpretable model for bivariate responses.

In future research, we shall consider multivariate hierarchical distributional regression models following the ideas of Lang *et al*. (2014). In addition, the multivariate normal model, the multivariate probit model and combinations of binary and continuous covariates in a joint latent normal model will be studied in detail to assess their potential as alternatives to well-known model types in economics such as Heckman's selection model. The fact that we provide both joint estimation of effects on several responses and allow for effects not only on the means can be expected to give rise to interesting new insight in corresponding applications.

Acknowledgements

We thank two referees and the Joint Editor for their careful review which was very helpful in improving on a first version of this paper. The work of Nadja Klein and Thomas Kneib was supported by the German Research Foundation via the research training group 1644 on 'Scaling problems in statistics' and research projects KN 922/4-1/2.

Appendix A: Full conditionals of latent variables in the bivariate probit model

In the probit model, the observable binary outcomes are replaced by latent variables as introduced in Section 2.1.1. For simplicity we describe the procedure for the example of a bivariate probit model where $\mathbf{y}^*$ is jointly bivariate normally distributed with expectation $\boldsymbol{\mu} = (\eta^{\mu_1}, \eta^{\mu_2})$ and covariance matrix

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

It follows that the full conditionals of $\mathbf{y}^*$ are truncated bivariate normal distributions

$$\mathbf{y}^*|\cdot \sim N_{[\mathbf{a},\mathbf{b}]}\,(\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

with

$$[\mathbf{a}, \mathbf{b}] = \begin{cases} [0,\infty) \times [0,\infty) & y_1 = 1, y_2 = 1, \\ [0,\infty) \times (-\infty,0] & y_1 = 1, y_2 = 0, \\ (-\infty,0] \times [0,\infty) & y_1 = 0, y_2 = 1, \\ (-\infty,0] \times (-\infty,0] & y_1 = 0, y_2 = 0. \end{cases}$$

Sampling from a truncated bivariate normal distribution has for example been studied by Robert (1995)

and a common way is to derive a Gibbs sampler for realizing random numbers from the truncated bivariate normal distribution desired. Although the Gibbs sampler usually converges in a small number of steps (see the appendix of Ambrosino *et al*. (2014)), we propose to draw the components of $\mathbf{y}^*$ separately from their conditional distributions. Although the marginal distributions are no longer truncated Gaussian, it can easily be shown that the conditional distributions follow truncated univariate normal distributions, i.e.

$$y_i^* | y_j^*, \cdot \sim \begin{cases} N_{[0,\infty)}\{\mu_i + \rho(y_j^* - \mu_j), (1-\rho^2)\} & \text{if } y_i = 1, \\ N_{(-\infty,0]}\{\mu_i + \rho(y_j^* - \mu_j), (1-\rho^2)\} & \text{if } y_i = 0 \end{cases}$$

for $i, j = 1, 2$ and $i \neq j$. We expect that, in the bivariate setting, sampling from the conditional distributions will in many cases be more efficient and also avoids the necessity to consider sampling from bivariate truncated normals. In higher dimensions, the computational overhead to determine the conditional distributions will typically outweigh these gains and therefore the Gibbs sampler may be the preferable option in these cases.

References

Albert, J. H. and Chib, S. (1993) Bayesian analysis of binary and polychotomous response data. *J. Am. Statist. Ass.*, **88**, 669–679.

Ambrosino, C., Chandler, R. E. and Todd, M. C. (2014) Rainfall-derived growing season characteristics for agricultural impact assessments in South Africa. *Theor. Appl. Clim.*, **115**, 411–426.

Belitz, C., Hübner, J., Klasen, S. and Lang, S. (2010) Determinants of the socioeconomic and spatial pattern of undernutrition by sex in India: a geoadditive semi-parametric regression approach. In *Statistical Modelling and Regression Structures—Festschrift in Honour of Ludwig Fahrmeir* (eds T. Kneib and G. Tutz), pp. 155–179. Heidelberg: Physica.

Belitz, C. and Lang, S. (2008) Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computnl Statist. Data Anal.*, **53**, 61–81.

Brezger, A. and Lang, S. (2006) Generalized structured additive regression based on Bayesian P-splines. *Computnl Statist. Data Anal.*, **50**, 967–991.

Bundesinstitut für Bau-, Stadt- und Raumforschung (2002) *Regierungsbezirke, GeoBasis-BKG*. Bonn: Bundesinstitut für Bau-, Stadt- und Raumforschung.

Caputo, A., Foraita, R., Klasen, S. and Pigeot, I. (2003) Undernutrition in Benin—an analysis based on graphical models. *Socl Sci. Med.*, **56**, 1677–1691.

Chakraborty, B. (2003) On multivariate quantile regression. *J. Statist. Planng Inf.*, **110**, 109–132.

Chaudhuri, P. (1996) On a geometric notion of quantiles for multivariate data. *J. Am. Statist. Ass.*, **91**, 862–872.

Chen, M. H. and Dey, D. K. (2000) Bayesian analysis for correlated ordinal data models. In *Generalized Linear Models: a Bayesian Perspective* (eds D. K. Dey, S. K. Ghosh and B. K. Mallick), pp. 133–159. New York: Dekker.

Dunn, P. K. and Smyth, G. K. (1996) Randomized quantile residuals. *J. Computnl Graph. Statist.*, **5**, 236–245.

Eilers, P. H. and Marx, B. D. (1996) Flexible smoothing using B-splines and penalized likelihood. *Statist. Sci.*, **11**, 89–121.

Fahrmeir, L. and Kneib, T. (2011) *Bayesian Smoothing and Regression for Longitudinal, Spatial and Event History Data*. New York: Oxford University Press.

Fahrmeir, L., Kneib, T., Lang, S. and Marx, B. (2013) *Regression—Models, Methods and Applications*. Berlin: Springer.

Fahrmeir, L. and Lang, S. (2001) Bayesian semiparametric regression analysis of multicategorical time-space data. *Ann. Inst. Statist. Math.*, **53**, 11–30.

Fenske, N., Kneib, T. and Hothorn, T. (2011) Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression. *J. Am. Statist. Ass.*, **106**, 494–510.

Frühwirth-Schnatter, S., Frühwirth, R., Held, L. and Rue, H. (2009) Improved auxiliary mixture sampling for hierarchical models of non-Gaussian data. *Statist. Comput.*, **19**, 479–492.

Gamerman, D. (1997) Sampling from the posterior distribution in generalized linear mixed models. *Statist. Comput.*, **7**, 57–68.

Gneiting, T. and Raftery, A. E. (2007) Strictly proper scoring rules, prediction, and estimation. *J. Am. Statist. Ass.*, **102**, 359–378.

Greene, W. H. (2011) *Econometric Analysis*, 7th edn. Harlow: Prentice–Hall.

Hastie, T. J. and Tibshirani, R. J. (1990) *Generalized Additive Models*. Boca Raton: Chapman and Hall–CRC.

Heckman, J. (1978) Dummy endogenous variables in a simultaneous equation system. *Econometrica*, **46**, 931–959.

Imai, K. and van Dyk, K. I. (2005) A Bayesian analysis of the multinomial probit model using marginal data augmentation. *J. Econmetr.*, **124**, 311–334.

Klasen, S. (2008) Poverty, undernutrition, and child mortality: some inter-regional puzzles and their implications for research and policy. *J. Econ. Inequalty*, **6**, 89–115.
Klasen, S. and Moradi, A. (2000) The nutritional status of elites in India, Kenya, and Zambia: an appropriate guide for developing reference standards for undernutrition? *Discussion Paper 217*. Georg-August-Universität Göttingen, Göttingen. (Available from `http://epub.ub.uni-muenchen.de/view/subjects/160101.html`.)
Klein, N., Kneib, T. and Lang, S. (2013) Bayesian structured additive distributional regression. *Technical Report*. Georg-August-Universität Göttingen, Göttingen. (Available from `http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2013-23.pdf`.)
Kotz, S., Balakrishnan, N. and Johnson, N. L. (2005) *Continuous Multivariate Distributions*, vol. 1, *Models and Applications*. New York: Wiley.
Krivobokova, T., Kneib, T. and Claeskens, G. (2010) Simultaneous confidence bands for penalized spline estimators. *J. Am. Statist. Ass.*, **105**, 852–863.
Lang, S., Adebayo, S. B., Fahrmeir, L. and Steiner, W. J. (2003) Bayesian geoadditive seemingly unrelated regression. *Computnl Statist.*, **18**, 263–292.
Lang, S. and Fahrmeir, L. (2001) Bayesian generalized additive mixed models: a simulation study. *Discussion Paper 230*. (Available from `http://www.uibk.ac.at/statistics/personal/lang/publications/`.)
Lang, S., Umlauf, N., Wechselberger, P., Harttgen, K. and Kneib, T. (2014) Multilevel structured additive regression. *Statist. Comput.*, **24**, 223–238.
Maddala, G. S. (1983) *Limited Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
McCullagh, P. and Nelder, J. A. (1989) *Generalized Linear Models*, 2nd edn. New York: Chapman and Hall.
Pourahmadi, M. (2011) Covariance estimation: the GLM and regularization perspectives. *Statist. Sci.*, **26**, 369–387.
Rigby, R. A. and Stasinopoulos, D. M. (2005) Generalized additive models for location, scale and shape (with discussion). *Appl. Statist.*, **54**, 507–554.
Robert, C. P. (1995) Simulation of truncated normal variables. *Statist. Comput.*, **5**, 121–125.
Rue, H. and Held, L. (2005) *Gaussian Markov Random Fields*. Boca Raton: Chapman and Hall–CRC.
Ruppert, D., Wand, M. P. and Carroll, R. J. (2003) *Semiparametric Regression*. New York: Cambridge University Press.
Scheipl, F., Fahrmeir, L. and Kneib, T. (2012) Spike-and-slab priors for function selection in structured additive regression models. *J. Am. Statist. Ass.*, **107**, 1518–1532.
Smith, M. and Kohn, R. (2000) Nonparametric seemingly unrelated regression. *J. Econmetr.*, **98**, 257–281.
Spiegelhalter, D. J., Best, N. G., Carlin, B. P. and van der Linde, A. (2002) Bayesian measures of model complexity and fit (with discussion). *J. R. Statist. Soc.* B, **64**, 583–639.
Tutz, G. (2011) *Regression for Categorical Data*. Cambridge: Cambridge University Press.
UNICEF (1998) *The State of the World's Children 1998: Focus on Nutrition*. Oxford: Oxford University Press.
Walter, F. (2010) Gelb oder Grün?: kleine Parteiengeschichte der besserverdienenden Mitte in Deutschland. *Transcript*.
Wiesenfarth, M., Krivobokova, T., Klasen, S. and Sperlich, S. (2012) Direct simultaneous inference in additive models and its application to model undernutrition. *J. Am. Statist. Ass.*, **107**, 1286–1296.
Winkelmann, R. (2008) *Econometric Analysis of Count Data*. Berlin: Springer.
Wood, S. N. (2006) *Generalized Additive Models: an Introduction with R*. Boca Raton: Chapman and Hall.
Woolridge, J. M. (2002) *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.
Zellner, A. (1962) An efficient method of estimating seemingly unrelated regression equations and tests for aggregation bias. *J. Am. Statist. Ass.*, **57**, 500–509.

Supporting information

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

'Bayesian structured additive distributional regression for multivariate responses supplement'.