Lang, Adebayo, Fahrmeir, Steiner:

Bayesian Geoadditive Seemingly Unrelated Regression

Sonderforschungsbereich 386, Paper 300 (2002)

Online unter: http://epub.ub.uni-muenchen.de/

Projektpartner

Bayesian Geoadditive Seemingly Unrelated Regression

Stefan Lang[1], Samson B. Adebayo[1], Ludwig Fahrmeir[1] and Winfried J. Steiner[2]

[1] Department of Statistics, University of Munich, Ludwigstr. 33,
80539 Munich, Germany.

[2] Department of Marketing, University of Regensburg, Universitätsstr. 31,
93053 Regensburg, Germany.

Abstract

Parametric seemingly unrelated regression (SUR) models are a common tool for multivariate regression analysis when error variables are reasonably correlated, so that separate univariate analysis may result in inefficient estimates of covariate effects.

A weakness of parametric models is that they require strong assumptions on the functional form of possibly nonlinear effects of metrical covariates. In this paper, we develop a Bayesian semiparametric SUR model, where the usual linear predictors are replaced by more flexible additive predictors allowing for simultaneous nonparametric estimation of such covariate effects and of spatial effects. The approach is based on appropriate smoothness priors which allow different forms and degrees of smoothness in a general framework. Inference is fully Bayesian and uses recent Markov chain Monte Carlo techniques.

Keywords Bayesian semiparametric models, correlated responses, Markov random fields, MCMC, P-splines.

1 Introduction

Multivariate regression analysis is needed in many applications. Neglecting existing association between response variables can lead to biased and inefficient estimation of covariate effects. Correspondingly, analysis of correlated response data has received great attention and interest. For multivariate Gaussian response, the parametric seemingly unrelated regression (SUR) model (Zellner, 1962) is a standard tool in econometrics. More recently, parametric SUR models have been developed for non-Gaussian responses, such as categorical or counted outcomes, see for example Chen and Dey (2000) and Winkelmann (2000) for recent works.

As in univariate regression, the assumption of a parametric linear predictor for assessing the impact of covariates on responses is often too restrictive in realistically complex situations. In particular, it is generally often difficult if not impossible to specify parametric functional forms for nonlinear effects of metrical covariates or of time scales in longitudinal studies in advance. For univariate responses, the development of more flexible non- and semiparametric regression models has been a main topic of recent research, see for example Hastie and Tibshirani (1990), Green and Silverman (1994), Fan and Gijbels (1996) or, for an introductory survey, Fahrmeir and Tutz (2001, ch. 5). There has also been substantial interest in extending parametric models for longitudinal or clustered data, where the same response variable is observed repeatedly, see Wild and Yee (1996), Lin and Carroll (2000) and Fahrmeir and Tutz (2001, ch. 6 & 7) for some recent works.

Astonishingly, there is a distinct lack of non- and semiparametric regression models for truly multivariate responses, in particular for seemingly unrelated regression. A notable exception is the Bayesian approach of Smith and Kohn (2000) to nonparametric seemingly unrelated regression. Their method uses basis function representations of unknown functions in combination with Bayesian variable selection and model averaging. In a sim-

ulation study, they show that the shape of nonlinear effects can be considerably biased and that estimation can become inefficient when applying separate univariate regressions instead of a multivariate model.

In this paper, we present a Bayesian approach to geoadditive seemingly unrelated regression (SUR), which is based on smoothness priors. Predictors incorporate linear parametric components, additive components for nonlinear effects of metrical covariates and a spatial component for geographical effects. The approach extends previous works of Fahrmeir and Lang (2001a) and Lang and Brezger (2002) for univariate generalized additive models or for multicategorical response (Fahrmeir and Lang, 2001b) to the present situation. Inference is fully Bayesian and relies on efficient MCMC techniques. The method is implemented in *BayesX*, an open domain software which can be downloaded from `http://www.stat.uni-muenchen.de/~lang/bayesx/bayesx.html`.

It is interesting to ask how much is gained by using SUR models instead of separate univariate regressions. For linear SUR models (Greene, 1993) it is well known that the greater the correlation of the errors, the greater the efficiency gain when using SUR, and the less correlation there is between the design matrices, the greater the gain. On the other side, if the equations are actually uncorrelated or if the design matrices are identical in all equations, SUR and ordinary least squares (OLS) regressions give the same results. It is to be expected that this is similar for additive and geoadditive SUR models. We study and illustrate this with applications to artificial and real data sets.

The rest of the paper is organized as follows: In Section 2 we describe our Bayesian semiparametric seemingly unrelated regression model. Section 3 outlines the MCMC procedure used for estimation. We demonstrate in Section 4 the usefulness of our approach through two simulation studies. Section 5 concludes this work with applications to marketing research and malnutrition of children in a developing country.

2 Bayesian geoadditive SUR

As common in hierarchical Bayesian models, we first describe the observation model. This is supplemented by appropriate prior assumptions for unknown parameters in a second stage.

2.1 Observation model

Suppose that regression data consist of observations $y_i = (y_{i1}, ..., y_{ik})'$, $i = 1, \ldots, n$, on a multivariate response y and on covariates. We distinguish between a vector $x_{ir} = (x_{ir1}, ..., x_{irp_r})'$ of *metrical* or *spatial* covariates whose influence on the r th component of y_i, will be modelled nonparametrically, and a further vector $v_{ir} = (v_{ir1}, ..., v_{irq_r})'$ of covariates, whose effect is modelled in the common usual form. We call a covariate *spatial* if it provides information in which region of a geographical map a particular observation has been made. For each component y_{ir}, $r = 1, ..., k$, of the response we assume a semiparametric regression model

$$y_{ir} = \eta_{ir} + \varepsilon_{ir}, \quad i = 1, \ldots, n, \tag{1}$$

with additive predictors

$$\eta_{ir} = f_{r1}(x_{ir1}) + ... + f_{rp_r}(x_{irp_r}) + v_{ir}'\gamma_r. \tag{2}$$

The functions f_{rj} are possibly nonlinear functions of metrical or spatial covariates. Type and degree of smoothness is controlled by priors described in the following section. The linear combination $v_{ir}'\gamma_r$ corresponds to the usual parametric part of the predictor, including an intercept term. Note that the mean levels of the unknown functions are not identifiable. To ensure identifiability, the functions are constrained to have zero means. The errors $\varepsilon_i = (\varepsilon_{i1}, \ldots, \varepsilon_{ik})$, $i = 1, \ldots, n$, are assumed to be i.i.d. multivariate Gaussian with mean zero and a covariance matrix Σ, i.e. $\varepsilon_i|\Sigma \sim N(0, \Sigma)$. This implies with

$\eta_i = (\eta_{i1}, ..., \eta_{ik})'$ that

$$y_i|\eta_i, \Sigma \sim N(\eta_i, \Sigma), \tag{3}$$

where responses y_i are conditionally independent, given the predictors η_i.

2.2 Prior specifications

For Bayesian inference, the unknown functions f_{rj}, the "fixed effects" parameters γ_r and the covariance matrix Σ of the errors are considered as random variables and have to be supplemented by appropriate prior distributions.

2.2.1 Priors for nonlinear functions

We start by describing the *general form* of a prior for an unknown (possibly nonlinear) function f_{rj} of covariate x_{rj}. For notational convenience, we omit indices and illustrate the approach for a specific function f with covariate x. Let $f = (f(1), \ldots, f(n))'$ be the vector of corresponding function evaluations at the observed values of x. We express the vector f as the matrix product of a (deterministic, non random) design matrix X and a vector of unknown regression parameters β, i.e.

$$f = X\beta. \tag{4}$$

The general form of the prior for β is

$$\beta|\tau^2 \propto \exp(-\frac{1}{2\tau^2}\beta' K \beta) \tag{5}$$

where K is a *penalty matrix* that penalizes too abrupt jumps between neighbouring parameters. In most cases K will be rank deficient and therefore the prior for β improper. This implies that $\beta|\tau^2$ follows a partially improper Gaussian prior $\beta|\tau^2 \sim N(0, \tau^2 K^-)$ where K^- is a generalized inverse of the penalty matrix K.

The variance parameter τ^2 is the equivalent to the smoothing parameter in a frequentist approach and controls the trade off between flexibility and smoothness. In order to be able to estimate the "smoothing parameter" τ^2 simultaneously with β, a highly dispersed but proper hyperprior is assigned to it. We choose an inverse gamma distribution with hyperparameters a and b, i.e. $\tau^2 \sim IG(a, b)$. A prior for a function f is thus defined by specifying a design matrix X, a smoothness prior for β, and the hyperparameters a and b of the inverse gamma prior for τ^2. A particular prior depends on the *type of the covariate* and on *prior beliefs about smoothness of* f. We will now give specific examples.

Metrical covariates

Let us first consider the case of a metrical covariate x. Several alternatives are currently available for a smoothness prior of the unknown function f. Among others, these are random walk priors (Fahrmeir and Lang, 2001a), Bayesian smoothing splines (Hastie and Tisbshirani, 2000) and Bayesian P-splines (Lang and Brezger, 2002). In the following we will focus on P-splines. Compared to smoothing splines, a P-splines approach allows a more parsimonious parameterization, which is a particular advantage in a Bayesian approach where inference is based on MCMC techniques.

The basic assumption behind the P-splines approach is that the unknown smooth function f can be approximated by a spline of degree l defined on a set of equally spaced knots $\zeta_0 = x_{min} < \zeta_1 < \ldots < \zeta_{s-1} < \zeta_s = x_{max}$ within the domain of x. It is well known (de Boor, 1978) that such a spline can be written in terms of a linear combination of $m = s + l$ B-spline basis functions B_t, i.e.

$$f(x) = \sum_{t=1}^{m} \beta_t B_t(x).$$

A crucial point with splines is the choice of the number and also the position of the knots. For a small number of knots the resulting function space may be not flexible enough to capture the variability of the data. For a large number of knots estimated curves may

tend to overfit the data. As a remedy Eilers and Marx (1996) suggest a moderately large number of knots (usually between 20 and 40) to ensure enough flexibility, and to define a roughness penalty based on differences of adjacent regression coefficients to guarantee sufficient smoothness of the fitted curves. In our Bayesian approach we replace difference penalties by their stochastic analogues, i.e. first or second order random walk models for the regression coefficients

$$\beta_t = \beta_{t-1} + u_t, \quad \text{or} \quad \beta_t = 2\beta_{t-1} - \beta_{t-2} + u_t$$

with Gaussian errors $u_t \sim N(0, \tau^2)$ and diffuse priors $\beta_1 \propto const$, or β_1 and $\beta_2 \propto const$, for initial values, respectively. A first order random walk penalizes abrupt jumps $\beta_t - \beta_{t-1}$ between successive states and a second order random walk penalizes deviations from the linear trend $2\beta_{t-1} - \beta_{t-2}$. Obviously, a P-spline prior for f is of the general form (4) and (5). The columns of the design matrix X consist of the basis functions B_t evaluated at the observations. For example for a second order random walk prior we obtain

$$K = \begin{pmatrix}
1 & -2 & 1 & & & & & & \\
-2 & 5 & -4 & 1 & & & & & \\
1 & -4 & 6 & -4 & 1 & & & & \\
 & 1 & -4 & 6 & -4 & 1 & & & \\
 & & \ddots & \ddots & \ddots & \ddots & \ddots & & \\
 & & & 1 & -4 & 6 & -4 & 1 & \\
 & & & & 1 & -4 & 6 & -4 & 1 \\
 & & & & & 1 & -4 & 5 & -2 \\
 & & & & & & 1 & -2 & 1
\end{pmatrix}$$

for the penalty matrix in (5). Finally, we note that simple random walk priors as considered for function estimation in Fahrmeir and Lang (2001a, b) can be seen as a special case of Bayesian P-splines.

Spatial covariates

Suppose now that x is a spatial covariate, i.e. the values of x represent the location or site in connected geographical regions. For simplicity we assume that the regions are numbered consecutively, i.e. $x \in \{1, \ldots, S\}$. For the spatial effect of x we choose Markov random field priors common in spatial statistics. These priors reflect spatial neighbourhood relations. Usually one assumes that two sites s and j are neighbours if they share a common boundary although more sophisticated neighbourhood definitions are possible, see for example Besag et al. (1991). The most common Markov random field prior used is given by

$$\beta_s | \beta_j \; j \neq s, \tau^2 \sim N\left(\sum_{j \in \partial_s} \beta_j / N_s, \tau^2 / N_s\right), \tag{6}$$

where N_s is the number of adjacent sites and $j \in \partial_s$ denotes, that site j is a neighbour of site s. Thus the (conditional) mean of $f(s) = \beta_s$ is an unweighted average of function evaluations $f(j) = \beta_j$ of neighbouring sites j. Of course, generalizations of (6) using *weighted* averages for the conditional mean are possible, see Besag et al. (1991) and Fahrmeir and Lang (2001b) for an application. Since for every site one parameter is estimated, the design matrix X for a spatial effect is a simple 0/1 incidence matrix where the number of columns is equal to the number of sites. If observation i is located in site s then the element in the i th row and s th column of X is one, zero otherwise.

Further examples

P-splines and Markov random fields are not the only prior specifications supported by our approach. In fact, time varying seasonal effects as considered in Fahrmeir and Lang (2001a), 2 dimensional extensions of P-splines proposed in Lang and Brezger (2002) for modelling interactions of metrical covariates, or i.i.d random effects fit well in the general form (5) and are supported by our software. Details are omitted to keep the paper in reasonable length.

2.2.2 Further prior assumptions

In the absence of any prior knowledge a natural assumption for fixed effects parameters are independent diffuse priors, i.e.

$$p(\gamma_{rj}) \propto const, \quad r = 1, ..., k \quad j = 1, \ldots, q_r.$$

Another choice would be a multivariate Gaussian prior which allows to model prior knowledge.

For the covariance matrix Σ of the errors, we choose an inverse Wishart prior

$$\Sigma \sim IW(A, B) \tag{7}$$

where A is a scalar and B is a k x k symmetric and positive definite matrix. The p.d.f. is given by

$$P(\Sigma) \propto |\Sigma|^{-A-(k+1)/2} \exp(-tr(B\Sigma^{-1}))$$

for $|\Sigma| > 0$ and zero elsewhere. A standard choice for A is 1 and $B = diag(c, \ldots, c)$ with a small c, e.g. $c = 0.005$.

We complete the Bayesian model specification by the assumption that priors for function evaluations, fixed effects parameters and for variances are all mutually independent.

3 Bayesian Inference through MCMC

Inference is fully Bayesian and uses MCMC simulation, drawing from full conditionals of single parameters or blocks of parameters given the rest and the data. Some matrix notation will be introduced for deriving full conditionals. Let $y_{.r} = (y_{1r}, ..., y_{nr})'$ and $\eta_{.r} = (\eta_{1r}, ..., \eta_{nr})'$ denote the vector on the r th response variable and the corresponding vector of predictors. Then the additive predictors (1) can be written as

$$\eta_{.r} = \sum_{j=1}^{p_r} X_{rj}\beta_{rj} + V_r\gamma_r \tag{8}$$

where β_{rj} is the vector of regression parameters for function f_{rj} and X_{jr} is the respective design matrix. The matrix V_r is the usual design matrix for fixed effects with rows v'_{ir}, i=1,...,n.

Let $\beta = (..., \beta'_{rj}, ...)'$ denote the stacked vector of all regression parameters, $\tau^2 = (..., \tau^2_{rj}, ...)'$ the vector of corresponding variances τ^2_{rj} and $\gamma = (\gamma'_1, ..., \gamma'_r)'$ the stacked vector of all fixed effects parameters. Posterior analysis is then based on

$$p(\beta, \tau, \gamma, \Sigma|y) \;\propto\; \prod_{i=1}^{n} p(y_i|\eta_i, \Sigma) \prod_{r=1}^{k} \prod_{j=1}^{p_r} \Big(p(\beta_{rj}|\tau^2_{rj}) p(\tau^2_{rj}) \Big)\, p(\Sigma), \tag{9}$$

where $p(y_i|\eta_i)$ is given by the Gaussian observation model (3) for observation $y_i = (y_{i1}, ..., y_{ik})'$ and predictor η_i.

MCMC simulation is carried out by drawing from full conditionals for the blocks

$$\begin{aligned}
&\beta_{rj}, \quad r = 1, \ldots, k, \; j = 1, \ldots, p_r, \\
&\gamma_r, \quad r = 1, \ldots, k, \\
&\tau^2_{rj}, \quad r = 1, \ldots, k, j = 1, \ldots, p_r, \\
&\Sigma.
\end{aligned}$$

They are given as follows:

(i) The full conditional for β_{rj} is Gaussian, $\beta_{rj}|\cdot \;\sim\; N(\mu_{rj},\; P^{-1}_{rj})$, with precision matrix

$$P_{rj} = \frac{X'_{rj} X_{rj}}{\sigma^2_{r|-r}} \;+\; \frac{K_{rj}}{\tau^2_{rj}} \tag{10}$$

and mean

$$\mu_{rj} \;=\; P^{-1}_{rj} \left(\frac{1}{\sigma^2_{r|-r}} X'_{rj} (y_{.r} - o_{.r}) \right). \tag{11}$$

In (10) and (11), $\sigma^2_{r|-r}$ is the (conditional) variance

$$\sigma^2_{r|-r} \;=\; \sigma^2_r - \Sigma_{r,-r} \Sigma^{-1}_r \Sigma'_{r,-r},$$

derived from partitioning Σ into

$$\Sigma = \begin{pmatrix} \sigma^2_r & \Sigma_{r,-r} \\ \Sigma'_{r,-r} & \Sigma_r \end{pmatrix},$$

(after reordering for the r th component of the error variable). The vector $o_{.r}$ in (11) is an offset vector. The i th component o_{ir} of the offset vector $o_{.r}$ is given by

$$o_{ir} \;=\; \Sigma_{r,-r}\Sigma_r^{-1}(y_{i,-r} - \eta_{i,-r}) + \tilde{\eta}_{ir}, \tag{12}$$

where $y_{i,-r}$ and $\eta_{i,-r}$ are obtained from y_i and η_i by omitting the r th components of y_i and η_i, respectively. The working predictor $\tilde{\eta}_{ir}$ is obtained from η_{ir} by deleting the j th effect f_{rj}.

(ii) The full conditional for γ_r is Gaussian, $\gamma_r|\cdot \;\sim\; N(\mu_{\gamma_r}, P_{\gamma_r}^{-1})$, with

$$P_{\gamma_r} \;=\; \frac{1}{\sigma^2_{r|-r}} V_r'V_r, \qquad \mu_{\gamma_r} \;=\; (V_r'V_r)^{-1}V_r'(y_{.r} - o_{.r}). \tag{13}$$

where the offset $o_{.r}$ is defined in (12) and $\tilde{\eta}_{ir}$ now obtained from η_{ir} by deleting the linear fixed effects term $v_{ir}'\gamma_r$.

(iii) Full conditionals for the variance parameters τ^2_{rj} are inverse Gamma distributions with parameters

$$a'_{rj} \;=\; a_{rj} \;+\; \frac{rank(K_{rj})}{2}\;, \qquad b'_{rj} = b_{rj} \;+\; \frac{1}{2}\beta'_{rj}K_{rj}\beta_{rj}. \tag{14}$$

(iv) The full conditional for Σ is an inverse Wishart distribution with parameters

$$A' \;=\; A \;+\; \frac{n}{2}\;, \qquad B' \;=\; B \;+\; \frac{1}{2}\sum_{i=1}^{n}(y_i \;-\; \eta_i)(y_i \;-\; \eta_i)'. \tag{15}$$

All full conditionals involved have known distributions, hence Gibbs sampling can be used to update the parameters of the model.

A fast implementation of MCMC updates requires efficient sampling from the full conditionals for the regression parameters β_{rj} of nonlinear functions. Following Rue (2001) drawing random numbers from $p(\beta_{rj}|\cdot)$ is as follows:

(i) Compute the Cholesky decomposition $P_{rj} = LL'$ of the posterior precision matrix.

(ii) Solve $L'\beta_{rj} = z$, where z is a vector of independent standard Gaussians. It follows that $\beta_{rj} \sim N(0, P_{rj}^{-1})$.

(iii) Compute the mean μ_{rj} by solving

$$P_{rj}\mu_{rj} = \frac{1}{\sigma^2_{r|-r}} X'_{rj}(y_{.r} - o_{.r})$$

with respect to μ_{rj}. This is achieved by first solving by forward substitution $L\nu = \frac{1}{\sigma^2_{r|-r}} X'_{rj}(y_{.r} - o_{.r})$ followed by backward substitution $L'\mu_{rj} = \nu$.

(iv) Add μ_{rj} to the previously simulated β_{rj}, then $\beta_{rj} \sim N(\mu_{rj}, P_{rj}^{-1})$.

The algorithms involved take advantage of the special structure of the precision matrices. For P-splines the precision matrices are band matrices where the bandwidth is the maximum between the degree l of the spline and the order of the random walk. The precision matrices of spatial effects modelled by Markov random field priors are sparse matrices but usually no band matrices. However, the regions of a geographical map can be reordered according to the Cuthill Mc-Kee algorithm (see George and Liu (1981) p. 58 ff) to obtain band matrix like precision matrices. The bandsize of the precision matrix usually differs from row to row. Rue (2001) uses matrix operations for band matrices to draw random numbers from the high dimensional full conditionals, i.e the different band sizes in every row are not utilized. In our implementation the different band sizes are exploited by using the *envelope method* for Cholesky decompositions of sparse matrices as described in George and Liu (1981). Our limited experience shows that the speed of the computations improves up to 30%.

4 Simulation Studies

4.1 Simulation study 1

To investigate how well our approach performs, we first carry out a simulation study based on a model used also by Smith and Kohn (2000) but does not include spatial components. The model includes the same four functions as in Smith and Kohn (2000) and is specified through

$$\begin{aligned} y_1 &= \sin(8\pi x_1) + \varepsilon_1 \\ y_2 &= [\phi(x_2; 0.2, 0.05) + \phi(x_2; 0.6, 0.2)]/4 + \varepsilon_2 \\ y_3 &= 1.5x_3 + \varepsilon_3 \\ y_4 &= cos(2\pi x_4) + \varepsilon_4. \end{aligned}$$

The covariate values are i.i.d. samples from $x_1 \sim U(0,1)$, $x_2 \sim U(0,1)$ and

$$\begin{pmatrix} x_3 \\ x_4 \end{pmatrix} \sim N\Bigg(\begin{pmatrix} 0.5 \\ 0.5 \end{pmatrix}, 0.3 \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix} \Bigg).$$

Because the covariance matrix Σ reported in Smith and Kohn (2000) turned out not to be positive definite, we chose

$$\Sigma = \begin{pmatrix} 1 & 0.7 & 0.6 & 0.9 \\ & 1 & 0.7 & 0.9 \\ & & 1 & 0.7 \\ & & & 1 \end{pmatrix}.$$

We simulated 250 replications of the model each with $n = 100$ observations. For each replication we estimated a Bayesian SUR model with cubic P-splines and both first and second order random walk penalties. For comparison, we additionally estimated univariate Gaussian regression models ignoring the correlations of the errors. To assess the dependence of results on the hyperparameters we estimated the models with three different choices for

the hyperparameters a and b of the variances τ_{r1}^2. We used $a = 1, b = 0.005$, $a = b = 0.001$ and $a = b = 0.0001$.

Figure 1 a) - h) shows posterior mean estimates of f_1-f_4 averaged over the 250 replications for both multivariate fits (left panels) and the corresponding separate univariate fits (right panels). The results shown are based on second order random walk penalties only. For first order random walks we get almost identical results, except for function f_3 where a small bias can be observed at the boundaries. Figure 2 displays boxplots of $\log(MSE)$ for multivariate and univariate fits where the empirical mean squared error MSE for a function f is defined as

$$MSE(\hat{f}) = \sqrt{\frac{1}{100}\sum_{i=1}^{100}(f(x_i) - \hat{f}(x_i))^2}.$$

Both Figures 1 and 2 present results only for the choice $a = b = 0.0001$ of hyperparameters. Figure 3 compares boxplots of $\log(MSE)$ for the three choices of hyperparameters. The figure shows results only for function f_1, for the functions $f_2 - f_4$ we obtain comparable results. Finally, Table 1 investigates the average coverage of pointwise credible intervals based on nominal levels of 80% and 95%. Using MCMC simulation techniques, credible intervals are estimated by computing the respective quantiles of the sampled function evaluations. From Figures 1-3 and Table 1 we can draw the following conclusions:

- Compared to separate univariate regressions, fitting a SUR model considerably reduces the estimation bias and the MSE. The differences become smaller for the linear function f_3.

- The dependence of results on the choice of the order of the penalty is very small for SUR regressions, but slightly (sometimes considerably) higher for univariate regressions (compare the last three boxplots of Figure 3 b).

- The dependence of results on different choices of hyperparameters is negligible for

SUR regressions. For univariate regressions the dependence is stronger although the results for $f_2 - f_4$ not shown in Figure 3 are less severe.

- Coverage rates for SUR regressions are generally closer to the respective nominal level than for the univariate regressions. The smallest differences are once again obtained for the linear function f_3.

We also simulated data with uncorrelated errors, i.e. $\Sigma = I$, to compare the SUR fit to separate regression fits in this situation. It turned out that there is no practical loss of efficiency when applying a SUR model in a situation where the errors are actually uncorrelated.

4.2 Simulation study 2: a geoadditive SUR model

In this study we investigate how well nonparametric and spatial components can be recovered and separated from each other in a geoadditive model. Simulations are based on the model

$$y_{i1} = f_{11}(x_{i1}) + f_{12}(s_{i1}) + \varepsilon_{i1}$$

$$y_{i2} = f_{21}(x_{i2}) + f_{22}(s_{i2}) + \varepsilon_{i2}$$

Here, x_1 and x_2 are drawn again from $U(0,1)$ and f_{11} and f_{21} are identical to the functions f_1 and f_2 used in the first simulation study. $s = (s_1, s_2)$ are centroids of districts in a map of Zambia and $f_{12}(s)$ and $f_{22}(s)$ are bivariate functions of the centroids shown in Figure 4 a) and b). Note that we assume for every observation different districts in equation one and two. Several experiments with equal districts showed relatively little differences between SUR and univariate regressions, although we did not get identical results (e.g. because of different estimates for the variance parameters). We observed a tendency to slightly better results with SUR regressions for highly curved functions and to almost

identical results for linear functions. The error covariance matrix in our example was set to

$$\Sigma = \begin{pmatrix} 1 & 0.9 \\ 0.9 & 1 \end{pmatrix}.$$

As in Section 4.1, results are based on 250 simulation runs. The dependence on the hyperparameters is assessed with the same three choices for a and b as in 4.1.

Results for the functions f_{11} and f_{21} of the metrical covariates x_1 and x_2 are almost identical to Section 4.1 and therefore completely omitted. Figure 4 c) - f) shows posterior means for the functions f_{12} and f_{22} averaged over the 250 replications. The graphs c) and d) show results for SUR and the graphs e) and f) for univariate regressions. Results are restricted to the choice $a = b = 0.0001$ for the hyperparameters. Comparable results are obtained for the other choices. Figure 5 displays boxplots of $\log(MSE)$ for SUR and univariate regressions and for the three choices of hyperparameters a and b. Table 1 investigates the average coverage of pointwise credible intervals based on nominal levels of 80% and 95%. We draw the following conclusions:

- Fitting a SUR model reduces the estimation bias and the MSE only for function f_{12}. For the linear function f_{22} results are almost identical.

- Results for spatial functions are more dependent on the choice of hyperparameters as for metrical functions. Both for SUR and univariate regressions results are considerably improved by the choices $a = b = 0.001$ and $a = b = 0.0001$.

- Average coverages are close to the nominal levels for SUR with $a = 1, b = 0.005$ while for univariate regressions coverages are below the nominal level. For the choices $a = b = 0.001$ and $a = b = 0.0001$ average coverages are close to or above the nominal levels for both SUR and univariate regressions. For these choices and function f_{12}, average coverages are considerably above the nominal level for SUR.

5 Applications to real data

5.1 Sales of Orange Juice

Recently, semiparametric regression has received much attention in the marketing literature. E.g., van Heerde et al. (2001) proposed a kernel based approach to estimate the functional relationship between a brand's unit sales and price discounts while modeling other predictors parametrically. Similarly, Hruschka (2002) developed a semiparametric market share attraction model and used cubic smoothing splines to estimate price effects on a brand's market share. In empirical applications, both semiparametric models performed better compared to strictly parametric model specifications in terms of MSE (for estimation and validation samples) or BIC and bootstrapped error sum of squares. A Bayesian (but) parametric SUR sales response model to derive micro-marketing pricing strategies has been presented by Montgomery (1997).

This marketing application also deals with sales response modeling and serves as a teaching example, demonstrating some features of semiparametric SUR models. We apply our model to a data set which consists of weekly sales and prices for 6 brands (j=1,...,6) of the product category orange juice, collected over a 2-year time span (t=1,...,104 weeks, starting in January 2000) in 6 retail stores ($k = 1, \ldots, 6$). The data were provided by MADAKOM, 50825 Cologne, Germany.

To keep things simple for illustration purposes, our primary interest is in own-brand price effects. We therefore model a brand's unit sales as a function of its own price to determine the shape of the price response function, adjusting for calendar time. Let n_{jkt} denote the raw unit sales of brand j in store k and week t. To correct for skewness and for different store sizes, we transformed the raw unit sales to the working responses

$$y_{1kt} = \frac{\ln n_{1kt}}{1/104 \sum_t \ln n_{1kt}}, \cdots, y_{6kt} = \frac{\ln n_{6kt}}{1/104 \sum_t \ln n_{6kt}}.$$

The nominator adjusts for skewness, and the denominator for the different store sizes. We further pooled the data over the stores ($k = 1, \ldots, 6$). In a first attempt, we based our analyses on the working SUR model with 6 equations for the 6 brands

$$y_{jkt} = f_{j1}(price_{jkt}) + f_{j2}(t) + \varepsilon_{jkt}, \quad j = 1, \cdots, 6,$$

using cubic P-splines and second order random walk priors for the smooth functions. The estimated 6×6 correlation matrix for the errors is estimated as

$$\hat{\Sigma} = \begin{pmatrix} 1.00 & 0.45 & 0.46 & 0.14 & 0.22 & 0.17 \\ 0.45 & 1.00 & 0.44 & 0.22 & 0.33 & 0.20 \\ 0.46 & 0.44 & 1.00 & 0.10 & 0.22 & 0.19 \\ 0.14 & 0.22 & 0.10 & 1.00 & 0.28 & 0.27 \\ 0.22 & 0.33 & 0.22 & 0.28 & 1.00 & 0.29 \\ 0.17 & 0.20 & 0.19 & 0.27 & 0.29 & 1.00 \end{pmatrix}$$

showing moderate correlations. Figure 6 displays the 6 own-price response functions, estimated with the SUR model. For low to medium prices, the functions show the decreasing pattern one would expect. For higher prices, however, some curves indicate a slight increase in brand sales toward the upper limit of the observed brand prices. Clearly, this nonmonotonicity would be hardly interpretable. Interestingly, 4 out of 6 curves are of a reverse-S shape indicating saturation and threshold levels. Figure 7 picks out the effects of calendar time for two brands. There is a clear seasonal pattern, with a higher effect during winter time.

Figure 8 shows price response functions for two selected brands, now estimated by univariate regressions. Compared to the results from the SUR model, the overall shapes remain similar, but the curves are rougher and look less robust. This is caused by the lack of information on the other brands' sales and prices that is not used in these univariate, separate regressions, respectively.

We finally illustrate a typical issue in SUR modelling: Frequently, correlation of the errors is caused by unobserved heterogeneity or omitted regressors. To demonstrate this aspect, we extended our working SUR model by including additional store dummies (which we deliberately omitted in the starting model) in each of the six brand equations to capture store-specific effects. Figure 9 shows the resulting price response curves for two selected brands from this revised SUR model. As a result of incorporating the store indicator variables, the price response functions now show the expected monotonically decreasing pattern. Comparing the estimated covariance matrices for both SUR models, we also notice much smaller correlations in the extended SUR model.

5.2 Child Malnutrition in Zambia

In the second application, errors are reasonably correlated but the design matrices are very close to each other, so that we cannot expect much gain in efficiency by using a geoadditive SUR model. We give some further comments in the conclusion. Child malnutrition is a problem of great social and economic relevance in many developing countries. We will consider the influence of socio-demographic variables as well as district-specific regional effects on two anthropometric indices: *stunting*- which is insufficient height-for-age (an indication of chronic undernutrition) and *underweight*- which is insufficient weight-for-height. The indices are defined as standard deviation units (z-scores) from the median of a reference population. The z-scores for child i with anthropometric index AI_i are defined as $z_i = (AI_i - MAI)/\sigma$, where MAI refers to the median of the reference population and σ refers to the standard deviation of the reference population.

We will analyze data for 4847 children from the 1992 DHS survey for Zambia, one of the poorest sub-saharan African countries. While Kandala *et al.* (2001) considered *stunting* as the only response variable, we analyze the correlated responses *stunting* and *underweight*

simultaneously with a geoadditive SUR model. We present results for a selected model, including the metrical covariates *Age* (child's age in months) and BMI (mother's body mass index) as well as the spatial covariate S (districts in Zambia). In addition, we included socio-demographic categorical covariates such as household size, mother's educational attainment, child's gender type, mother's working status as well as fever and cough as indicators of acute disease. Fever and cough turned out to be nonsignificant for stunting, while fever has a significant negative effect on underweight. The geoadditive SUR model used for the final analysis was

$$\begin{aligned} stunting &= f_{11}(Age) + f_{12}(BMI) + f^1_{spat}(s) + v_1'\gamma_1 + \varepsilon_1 \\ underweight &= f_{21}(Age) + f_{22}(BMI) + f^2_{spat}(s) + v_2'\gamma_2 + \varepsilon_2, \end{aligned}$$

where v_2 contains fever and cough in addition to the categorical covariates in v_1. Thus the design matrices are almost the same in both equation, so that we cannot expect much gain in efficiency when using a SUR model. Nonlinear effects of *Age* and BMI are modelled by P-splines of degree 3 and second order random walk penalty, and spatial effects through a Markov random field prior.

The posterior mean estimate of the covariance matrix Σ, with correlation in the lower diagonal, is given by

$$\hat{\Sigma} = \begin{pmatrix} 0.805 & 0.519 \\ 0.656 & 0.777 \end{pmatrix}.$$

Figure 10 displays the nonlinear effects of child's age and mother's BMI on *stunting* and *underweight*, respectively. Shown are the posterior means within 80% and 95% pointwise credible intervals. For the age effects on both *stunting* and *underweight*, virtually similar patterns are noticeable. There is a continuous worsening of the nutritional status of children up till about 20 months of age. Such an immediate deterioration in nutritional status is worrisome as the worsening is associated with weaning age at around 4-6months. One reason for this finding could be that, according to surveys, most parents give their children

liquids other than breastmilk shortly after birth which might contribute to infections at the early age. The lower panels of Figure 10 reveals that influence of mother's BMI on *stunting* and *underweight* is approximately of an inverse U shape. This appears quite reasonable as obesity of the mother (possibly due to poor quality of diet) is of less risk for the nutritional status of the child.

District-specific regional effects are shown in Figure 11 through maps of significance (80%), showing regions with positively significant (white colored), negatively significant (black colored) and non significant (grey colored) regional effects. There is a sizeable difference between significantly worse undernutrition in the Central province and better nutrition in the Northern and South-Western provinces. Also similar patterns exist for *stunting* and *underweight*.

We do not display posterior estimates of the fixed effects γ_1 and γ_2 here. They are coherent with previous findings.

6 Conclusions

Simulations and applications illustrated that there can be considerable gain in using semiparametric SUR models instead of semiparametric regressions when errors are reasonably correlated and design matrices differ between equations. On the other hand, we do not have to expect much gain for situations as in the last application. We re-run this application with separate univariate regressions, and indeed there was not a big difference. It seems that correlation between errors may affect estimation of smoothing parameters, similarly as in nonparametric regressions with serially correlated errors but further investigation of this issue is necessary.

Another comment concerns specification and parameterization of the error covariance matrix. Using an inverse Wishart prior gives satisfactory results when the dimension of

the response vector is not too large ($k \leq 10$). For higher dimension more parsimonious parameterizations, for example based on Cholesky decompositions of the precision matrix (the inverse of the covariance matrix) will be a promising alternative (see Pourahmadi (1999) and Smith and Kohn (2002)). We also plan to extend our approach to SUR models for categorical responses, using Gaussian SUR models in latent threshold mechanisms as in Fahrmeir and Lang (2001b).

As pointed out by the referees, other MCMC updating schemes might be useful. In particular, joint block updates as in Knorr-Held and Rue (2002) could be considered, and the fixed effects γ could be integrated out analytically, similarly as in Gamerman et al. (2002). Note, however, that we had no problems with convergence or the mixing of the chains in both the simulation studies and the applications.

References

[1] Besag, J, York, Y. & Mollie, A. (1991): Bayesian Image Restoration with two Applications in Spatial Statistics (with discussion). *Annals of the Institute of Statistical Mathematics*, 43:1-59.

[2] Central Statistical Office [Zambia] and Ministry of Health and Macro International Inc. (1997): Zambia Demographic and Health Survey, 1996. *Calverton, Maryland: Central Statistical Office and Macro International Inc.*

[3] Chen, M. & Dey, D. (2000): Bayesian analysis for correlated Ordinal Data Models. *In D. Dey, S. Ghosh and B. Mallick (Eds)*, Generalized Linear Models. A Bayesian Perspective, pp133-159. *Marcel Dekker, New York.*

[4] De Boor, C. (1978): A Practical Guide to Splines. *Spriner-Verlag, New York.*

[5] Eilers, H.C. & Marx, D.B. (1996): Flexible Smoothing with B-Splines and Penalties. Statistical Science, 11(2):89-121.

[6] Fahrmeir, L. & Lang, S. (2001a): Bayesian Inference for generalized additive mixed models on Markov random field priors. *Applied Stat.*, 50(2):201-220.

[7] Fahrmeir, L. & Lang, S. (2001b): Bayesian Semiparametric Regression Analysis of Multicategorical Time-Space Data. *Annals of the Institute of Statistical Mathematics*, 52,No 1:1-18.

[8] Fahrmeir, L. & Tutz, G. (2001): Multivariate Statistical Modelling based on Generalized Linear Models (3rd ed), *Springer, New York.*

[9] Fan, J. & Gijbels, I. (1996): Local Polynomial Modelling and its Applications. *Chapman and Hall, London.*

[10] Gamerman, D., Moreira, A.R.B. & Rue, R. (2002): Space-varying regression models: specifications and simulations. *Computational Statistics and Data Analysis*, to appear.

[11] George, A. & Liu, J.W. (1981): Computer Solution of Large Sparse Positive Definite Systems. *Series in Computational Mathematics, Prentice-Hall.*

[12] Green, P.J. & Silverman, B.W. (1994): Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach. *Chapman and Hall, London.*

[13] Greene, W.H. (1993): *Econometric Analysis (2nd ed.)* Macmillian Publishing Co., New York.

[14] Hastie, T. & Tibshirani, R. (1990): Generalized Additive Models. *Chapman & Hall, London.*

[15] Hastie, T. & Tibshirani, R. (2000): Bayesian Backfitting. *Statistical Science*, 15:193-223.

[16] Hruschka, H. (2002): Market Share Analysis Using Semi-Parametric Attraction Models. *European Journal of Operational Research*, 138:212-225.

[17] Kandala, N.B., Lang, S., Klasen, S. & Fahrmeir, L. (2001): Semiparametric Analysis of the Socio-Demographic and Spatial Determinants of Undernutrition in Two African Countries. *Research in Official Statistics*, 1, 81-100.

[18] Knorr-Held, L. & Rue, H. (2002): On block updating in Markov random field models for disease mapping. *Scandinavian Journal of Statistics*, to appear.

[19] Lang, S. & Brezger, A. (2002): Bayesian P-Splines. *Journal of Computational and Graphical Statistics, to appear.*.

[20] Lin, X. & Carroll, R.J. (2000): Nonparametric Function Estimation for Clustered Data When the Predictor is Measured Without/With Error. *Journal of the American Statistical Association*, 95 (450):520-534.

[21] Montgomery, A.L. (1997): Creating Micro-Marketing Pricing Strategies Using Supermarket Scanner Data. *Marketing Science*, 16(4):315-337.

[22] Pourahmadi, M. (1999): Joint mean-covariance models with applications to longitudinal data: Unconstrained parameterisation. *Biometrika*, 86(3):677-690.

[23] Rue, H. (2001): Fast sampling of Gaussian Markov Random Fields with Applications. *Journal of the Royal Statistical Society B*, 63:325-338.

[24] Smith, M. & Kohn, R. (2000): Nonparametric Seemingly Unrelated Regression. Journal of Econometrics, 98:257-281.

[25] Smith, M. & Kohn, R. (2002): Bayesian Parsimonious Covariance Matrix Estimation. *Journal of the American Statistical Association*, to appear.

[26] Wild, C.J. & Yee, T.W. (1996): Additive Extensions to Generalized Estimating Equation Methods. *Journal of the Royal Statistical Society, Ser. B*, 58:711-725.

[27] Winkelmann, R. (2000): Seemingly Unrelated Negative Binomial. *Oxford Bulletin of Ecomics and Statistics*, 62(4): 553-560.

[28] van Heerde, H.J., Leeflang, P.S.H. & Wittink, D.R. (2001): Semiparametric Analysis to Estimate the Deal Effect Curve. *Journal of Marketing Research*, 38:197-215.

[29] Zellner, A. (1962): An Efficient Method of Estimating Seemingly Unrelated Regression Equations and Tests for Aggregation Bias. *Journal of the American Statistical Association*, 57:500-509.

Table 1: Simulation study 1 and 2. Average coverage rates in percent for the functions $f_1 - f_4$ of simulation study 1 and the spatial functions $_{12}$ and f_{22} of simulation study 2. The first column indicates the respective estimation technique used and the choice of hyperparameters. For simulation study 1 results are shown only for P-splines with second order random walk penalty.

	f_1	f_2	f_3	f_4	f_{12}	f_{22}
80% , SUR, a=1 b=0.005	80	76	83	80	82	79
80% , univ., a=1 b=0.005	36	64	85	73	71	79
95% , SUR, a=1 b=0.005	95	92	97	94	96	94
95% , univ, a=1 b=0.005	56	80	98	91	90	94
80% , SUR, a=b=0.001	80	77	83	79	87	81
80% , univ., a=b=0.001	56	69	85	78	80	82
95% , SUR, a=b=0.001	95	93	96	94	97	95
95% , univ, a=b=0.001	79	86	97	94	95	96
80% , SUR, a=b=0.0001	80	77	84	80	87	80
80% , univ., a=b=0.0001	54	68	85	78	79	82
95% , SUR, a=b=0.0001	95	93	97	94	95	95
95% , univ, a=b=0.0001	76	85	98	94	96	97

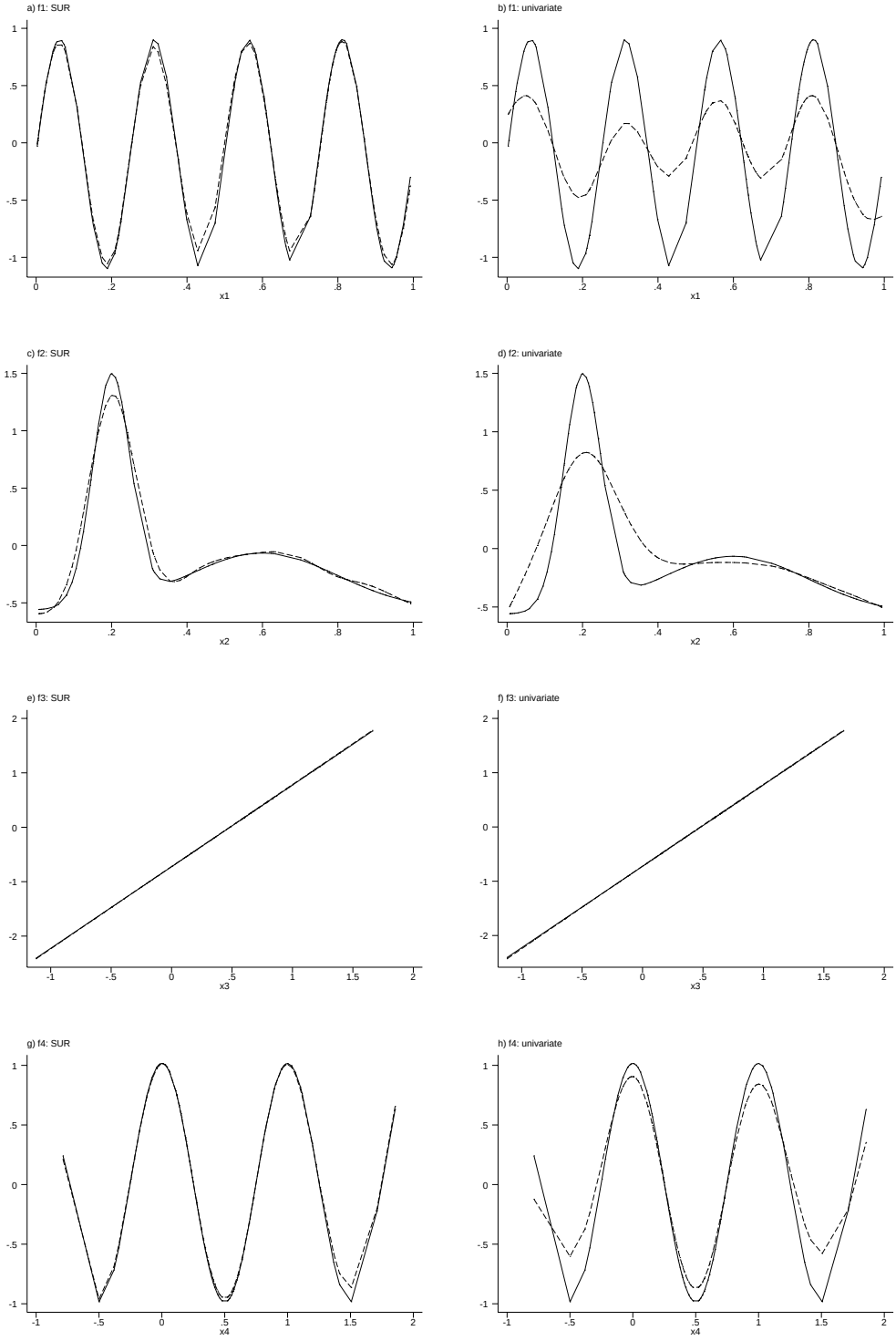

Figure 1: *Simulation study 1. Posterior mean estimates (dashed lines) of nonparametric functions f_1-f_4 averaged over the 250 replications. Left panels display SUR estimation results, right panels the corresponding univariate regressions ignoring correlations. For comparison the true functions are included (solid lines). The results are based on P-splines with second order random walk penalty and the choice $a = b = 0.0001$ for the hyperparameters of variances.*

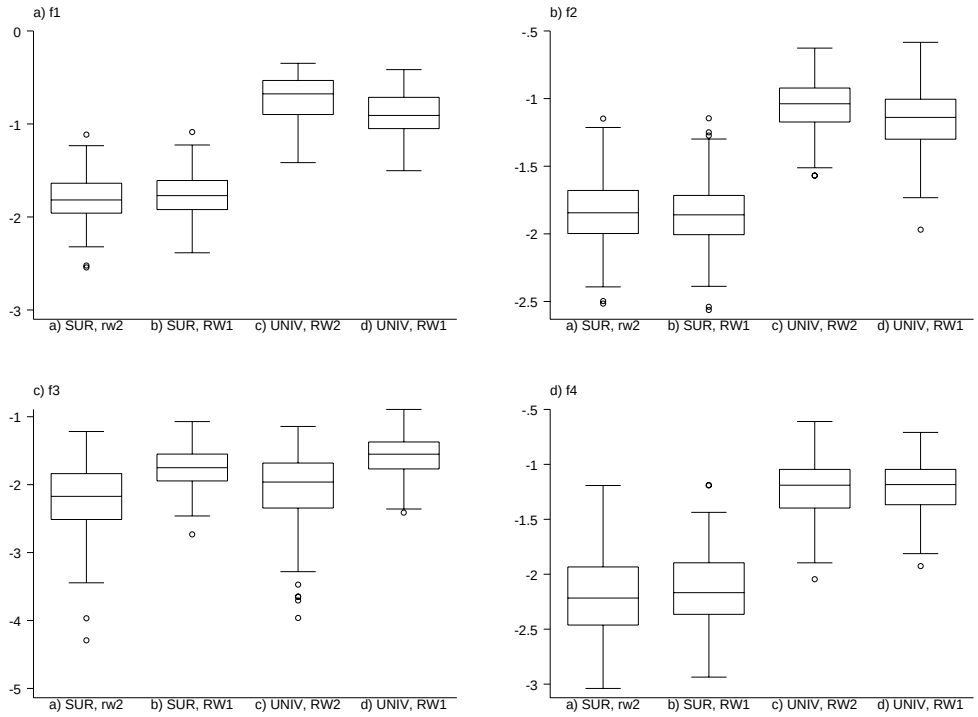

Figure 2: *Simulation study 1. Boxplots of* $log(MSE)$ *for SUR and univariate regressions, respectively. The results are based on the choice* $a = b = 0.0001$ *for the hyperparameters of variances.*

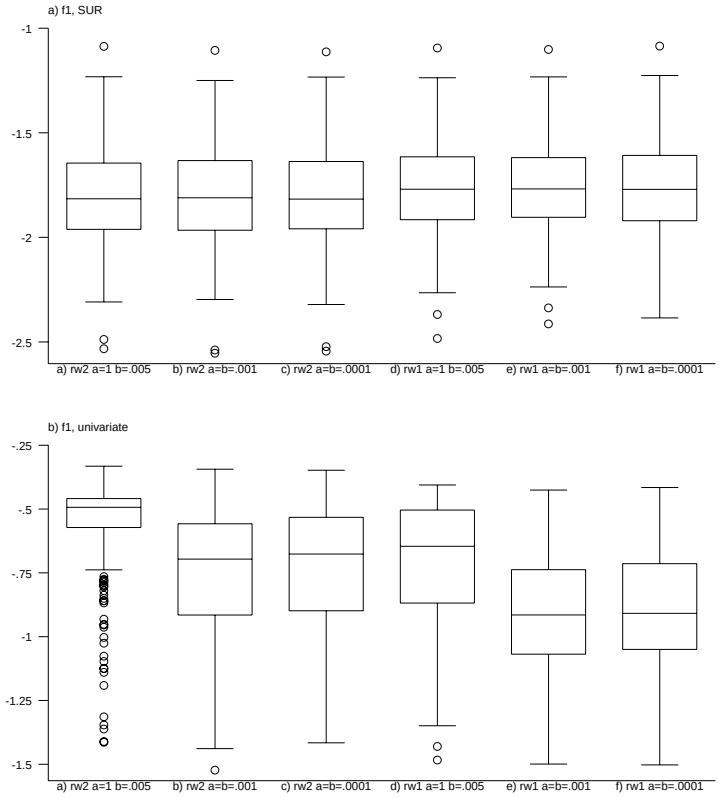

Figure 3: *Simulation study 1. Boxplots of* $log(MSE)$ *for the three different choices of the hyper-parameters* a *and* b *for function* f_1.

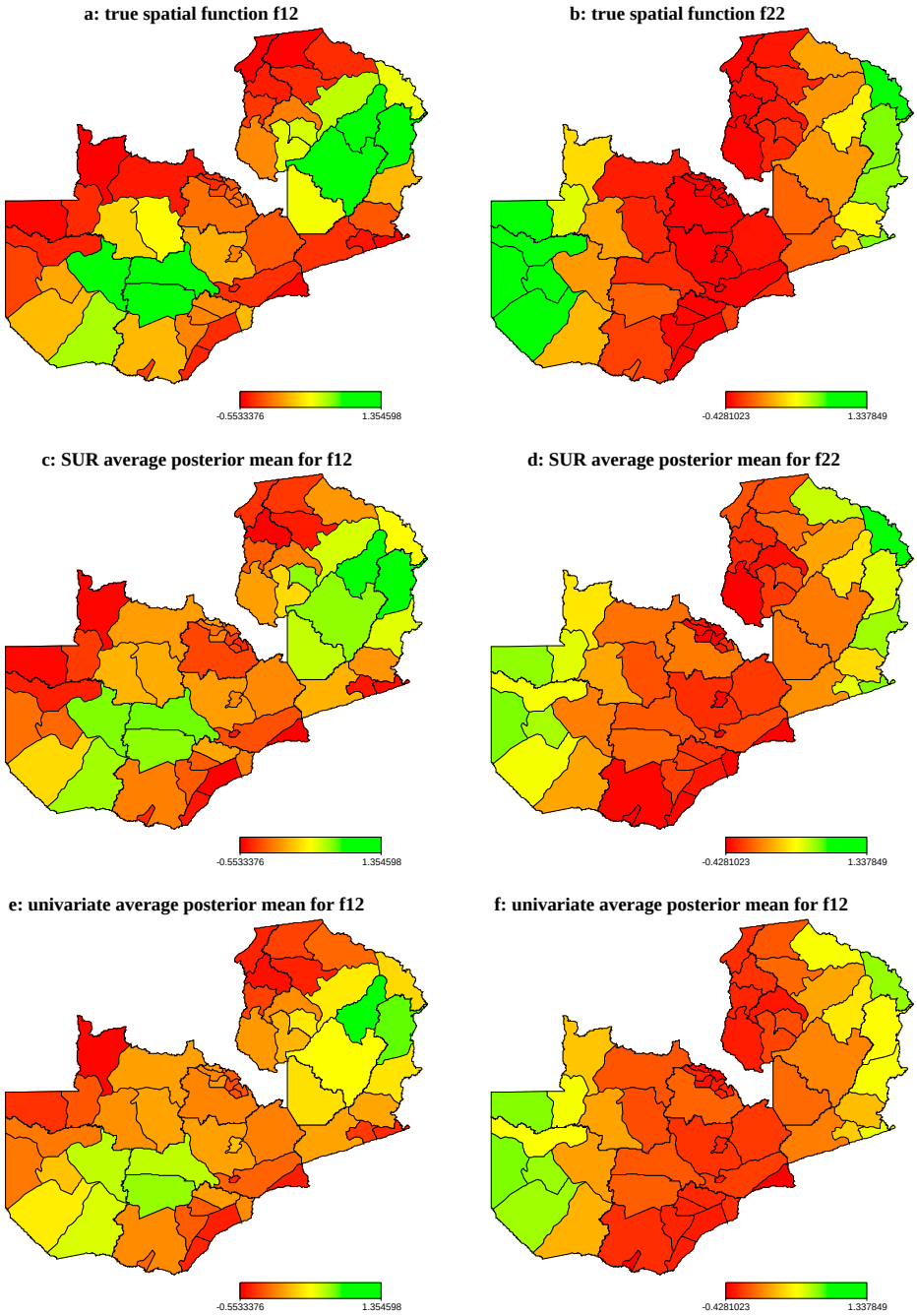

Figure 4: *Simulation study 2: The top graphs show maps of the true spatial effects for equation one (left panel) and equation two (right panel). The middle and bottom graphs show average posterior means of the spatial effects f_{12} (left panel) and f_{22} (right panel) for SUR and univariate regressions, respectively. The results presented are based on the choice $a = b = 0.0001$ for the hyperparameters of variances.*

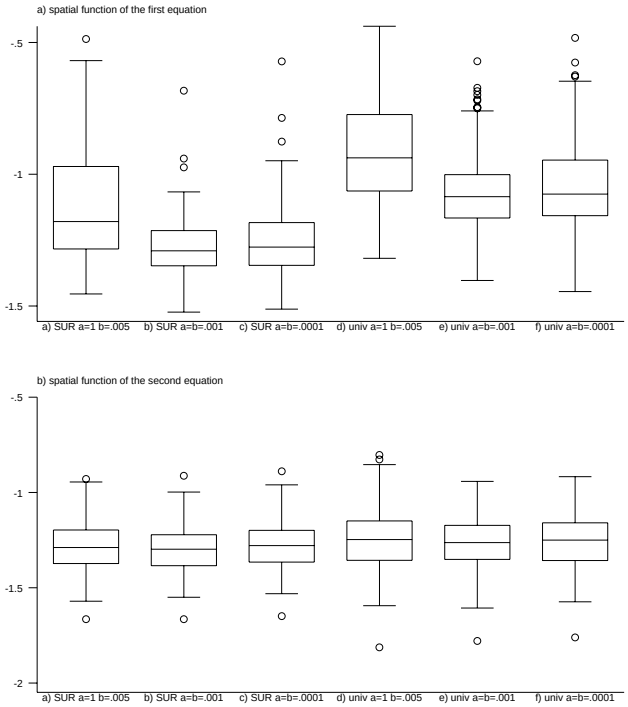

Figure 5: *Simulation study 2: Boxplots of* $log(MSE)$ *for SUR and univariate regressions, respectively.*

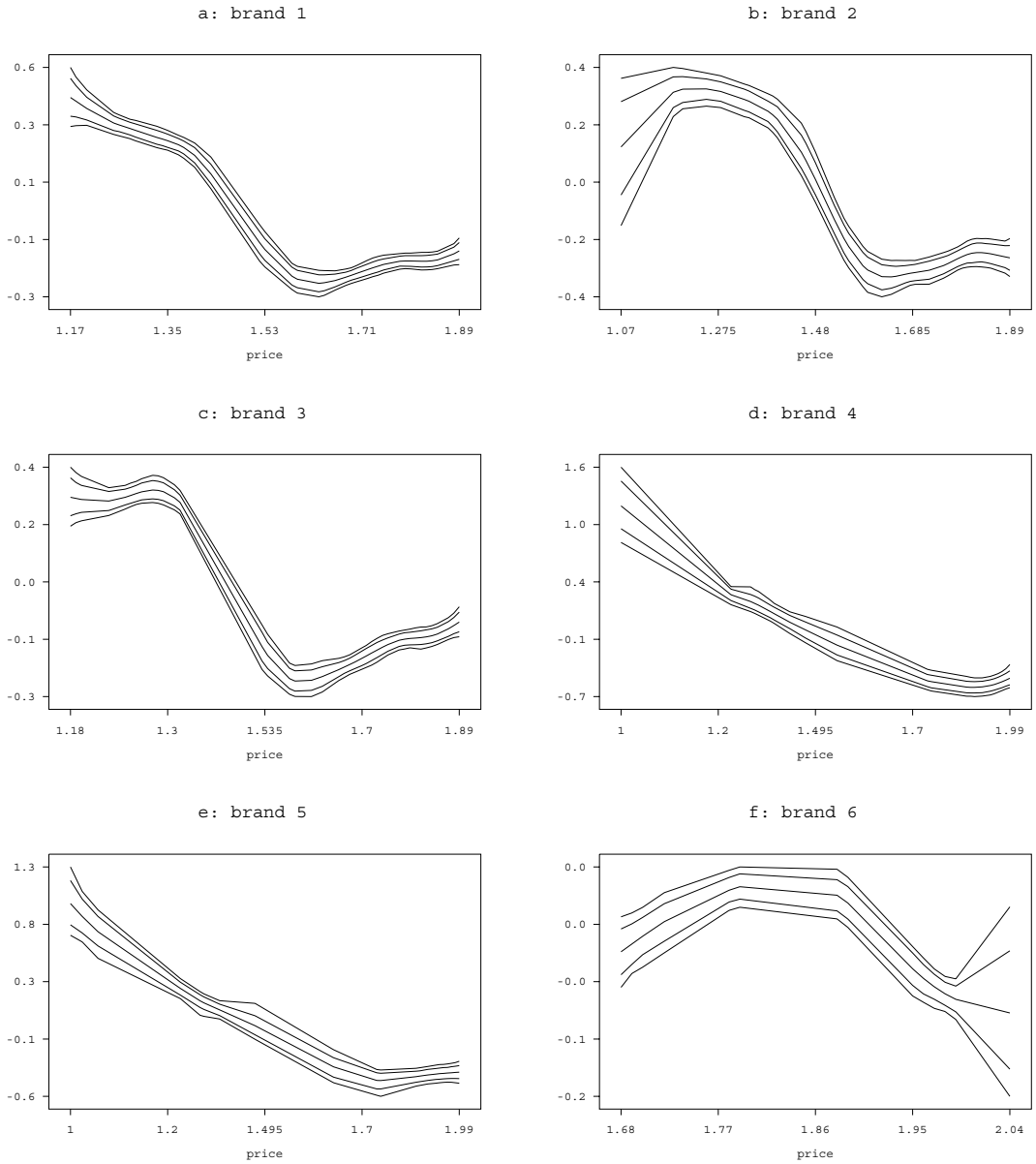

Figure 6: *Application on orange juice sales. Estimated price response functions for brands 1 to 6. Shown are the posterior means together with 80% and 95% pointwise credible intervals.*

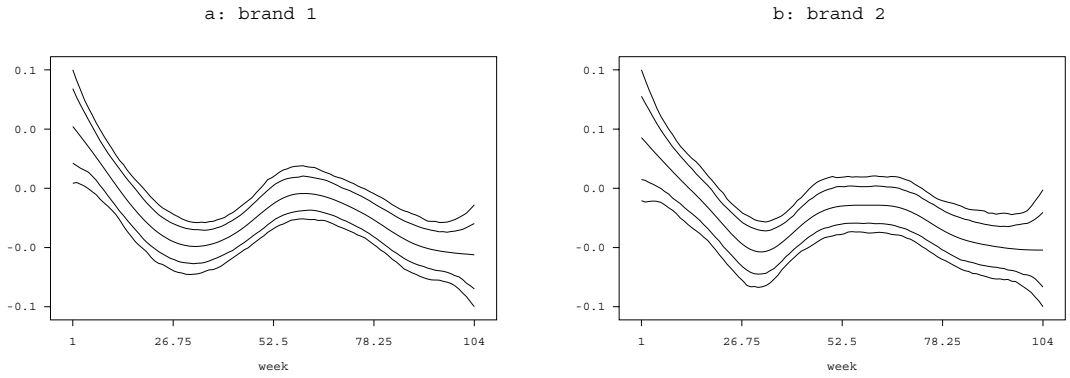

Figure 7: *Application on orange juice sales. Estimated time trends for brand 1 and 2. Shown are the posterior means together with 80% and 95% pointwise credible intervals.*

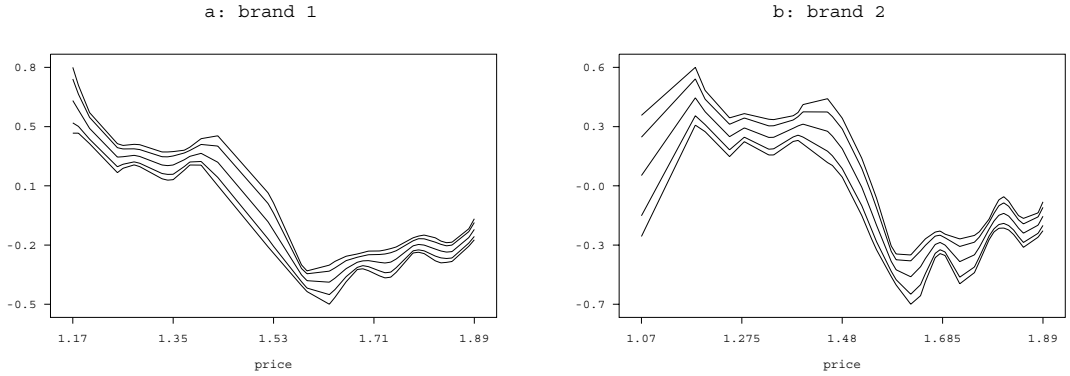

Figure 8: *Application on orange juice sales. Estimated price response functions for brand 1 and 2 based on separate univariate regressions ignoring correlations. Shown are the posterior means together with 80% and 95% pointwise credible intervals.*

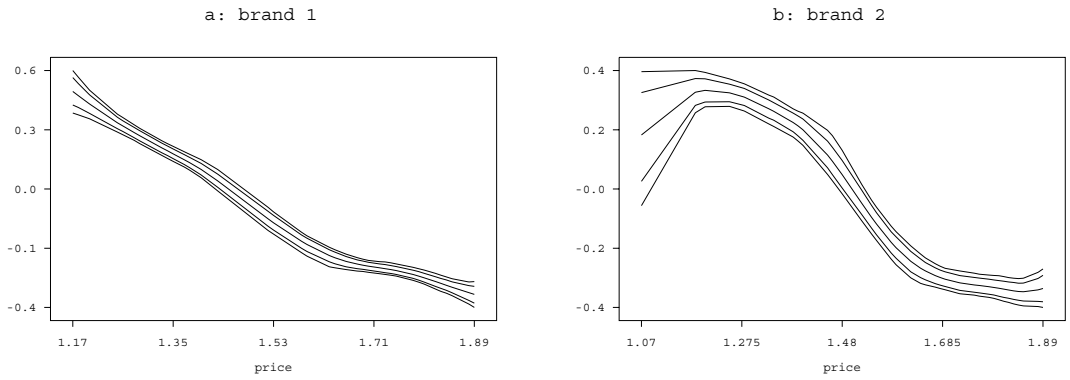

Figure 9: *Application on orange juice sales. Estimated price response functions for brand 1 and 2 when additional store dummies are included into the SUR model. Shown are the posterior means together with 80% and 95% pointwise credible intervals.*

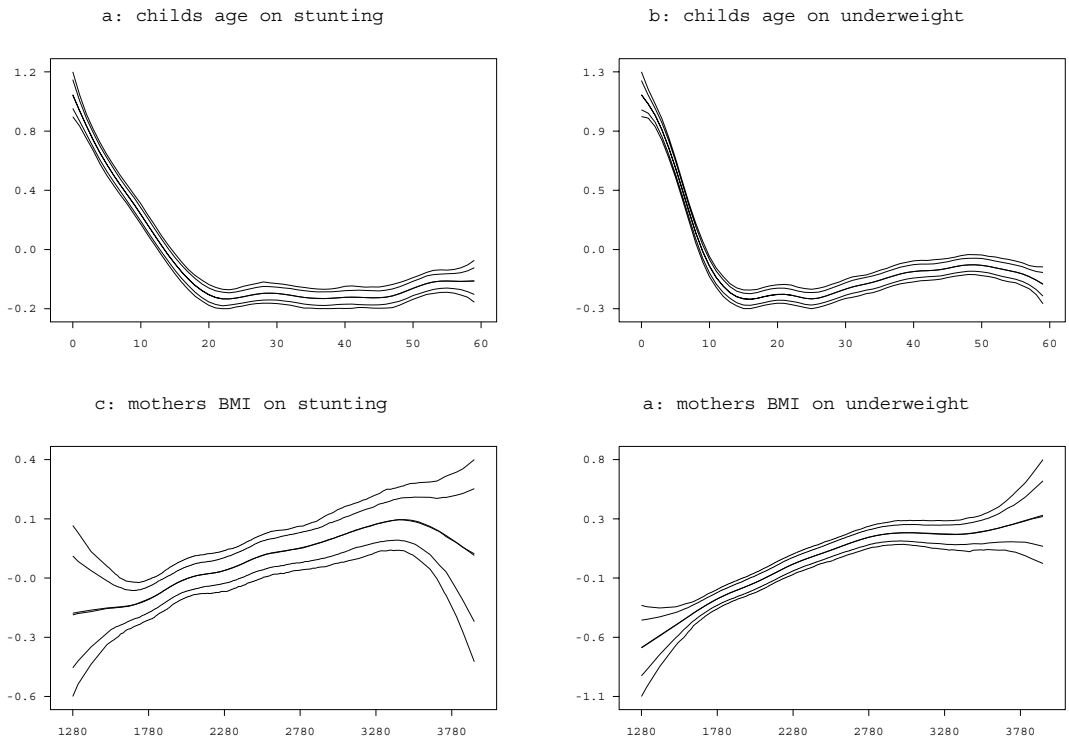

Figure 10: *Application on child malnutrition in Zambia. Nonparametric effects of child's age (top) and mother's BMI (bottom) on stunting (left panels) and underweight (right panels). Shown are the posterior means together with 80% and 95% pointwise credible intervals.*

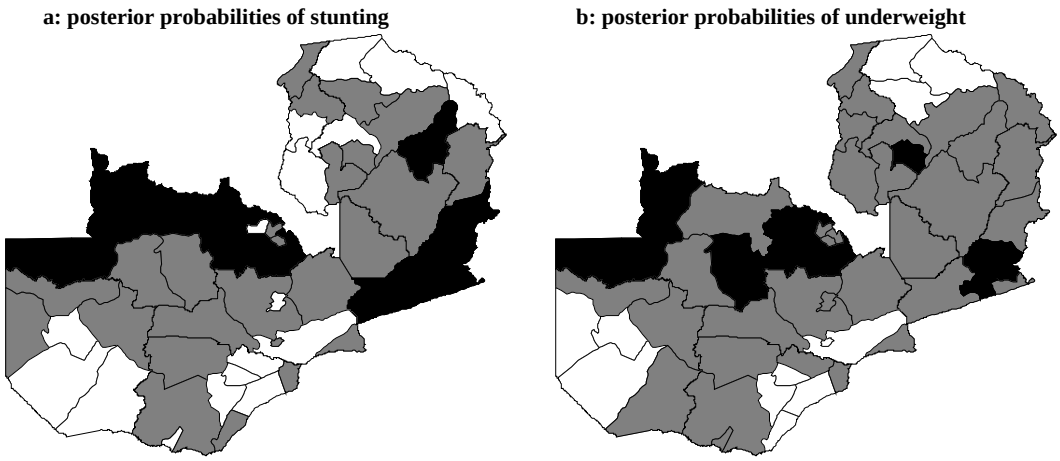

Figure 11: *Application on child malnutrition in Zambia. Posterior probabilities on stunting (left panel) and underweight (right panel) based on a nominal level of 80%.*