Estimation and Inference in Special Nonparametric Models with Applications to Topics in Development Economics

Dissertation
zur Erlangung des Doktorgrades
der Wirtschaftswissenschaftlichen Fakultät
der Georg-August-Universität Göttingen

vorgelegt von
Dipl. Stat. Manuel Wiesenfarth
geboren in Filderstadt

Göttingen, 2012

Erstgutachterin:	Prof. Dr. Tatyana Krivobokova
Zweitgutachter:	Prof. Stephan Klasen, Ph.D.
Tag der mündlichen Prüfung:	11. Mai 2012

Acknowledgements

I am most grateful and indebted to my principal advisor, Tatyana Krivobokova, for her patience, constant support and encouragement, for countless extensive discussions and for sharing her enormous knowledge. I can't put into words how much I have learned from her. Further, I would like to thank Stephan Klasen for initiating the research centre and the time for discussions which was much more than I could have ever expected given his tight schedule. He ultimately enabled me to get an insight into the exciting field of development economics. Furthermore, I am sincerely grateful to Thomas Kneib. I have been privileged to benefit from his support and guidance over many years.
I would like to extend my thanks to my officemates, my colleagues at the Courant Research Centre and at the Chair of Development Economics for the interesting discussions and good times I had during the last years in Göttingen. Lastly, I would like to thank my girlfriend Katrin for her loving support and patience and my parents for their unconditional support and encouragement throughout my academic career.

Table of Contents

1 Introduction **1**

1.1 Motivation and Objectives 1
1.2 A Case for Nonparametric Regression 3
1.3 Nonparametric Regression in a Nutshell 5
1.4 Outline 7

2 Direct Simultaneous Inference in Additive Models **10**

2.1 Introduction 10
2.2 Additive Models with Penalized Splines 13
2.3 Data on Childhood Undernutrition in Kenya 14
2.4 Simultaneous Confidence Bands 17
2.4.1 The Volume-of-tube Formula 17
2.4.2 Simultaneous Confidence Bands for Additive Models 18
2.4.3 Simultaneous Bands for Additive Models with Spatially Heterogeneous Components and Heteroscedastic Errors 19
2.5 A new Specification Test 22
2.6 Monte Carlo Studies 23
2.6.1 Simulation 1: Simultaneous Confidence Bands for Additive Models 23
2.6.2 Simulation 2: Additive Model with Locally Adaptive Smoothed Components and Heteroscedasticity 26
2.6.3 Simulation 3: Nonparametric Specification Test 27
2.7 Studying Undernutrition in Kenya 28
2.8 Discussion 31

3 Bayesian Geoadditive Sample Selection Models **34**

3.1 Introduction 34
3.2 Geoadditive Sample Selection Models 38
3.2.1 Parametric Effects 38
3.2.2 Nonparametric Effects 38
3.2.3 Varying Coefficient Terms 40

3.2.4 Spatial Effects . . . 40
3.2.5 Generic Model Representation . . . 41
3.2.6 Priors for the Error Term Covariance Matrix . . . 41
3.3 Bayesian Inference . . . 42
3.4 Simulations . . . 45
3.4.1 Simulation Study 1: Parametric Sample Selection Models . . . 45
3.4.2 Simulation Study 2: Geoadditive Sample Selection Models . . . 48
3.5 Relief Supply in Earthquake-Affected Communities in Pakistan . . . 53
3.5.1 Data Sources and Data Preparation . . . 53
3.5.2 Model and Prior Settings . . . 54
3.5.3 Results . . . 55
3.5.4 Discussion . . . 58
3.6 Outlook & Extensions . . . 59

4 Bayesian Nonparametric Instrumental Variable Regression 63
4.1 Introduction . . . 63
4.2 Additive Simultaneous Equations Model . . . 67
4.2.1 Parametric Effects . . . 68
4.2.2 Nonparametric Effects . . . 68
4.2.3 Joint Error Distribution . . . 70
4.2.4 Hyperparameter Choices . . . 73
4.3 Bayesian Inference . . . 74
4.3.1 Estimation . . . 74
4.3.2 Full Conditionals . . . 75
4.3.3 Smoothing Parameter Estimation . . . 78
4.3.4 Simultaneous Bayesian Credible Bands . . . 80
4.4 Simulations . . . 81
4.4.1 Parametric Model . . . 81
4.4.2 Nonparametric Model . . . 85
4.5 Application: Class Size Effects on Student Achievements . . . 89
4.6 Conclusion . . . 94

5 Software 95
5.1 Package AdaptFitOS . . . 95
5.1.1 Fitting a Non-adaptive Model . . . 95
5.1.2 Fitting a Model with Locally Adaptive Smoothing Parameters and Heteroscedastic Errors . . . 96

5.2 Package bayesIV 99
5.2.1 Bayesian Geoadditive Regression Models 100
5.2.2 Bayesian Geoadditive Sample Selection Models 100
5.2.3 Bayesian Nonparametric Instrumental Variable Regression 103

List of Figures

1.1 Examples of non-quadratic relationships in simulated data. 5

2.1 True functions in simulations 1 and 2 scaled to have variance 1. 24
2.2 Empirical power curves of the proposed test and RLRT test in simulation 3. 28
2.3 Estimated effects with corresponding partial residuals in model (2.5). . . . 29
2.4 Estimated effects in the final model. 30
2.5 Estimated error density, smoothing parameter function and residual standard deviation. 32
2.6 Estimated first derivative of the age effect and estimated age effect and its first derivative assuming that the recumbent length and standing height only differ by 0.3 cm. 33

3.1 Simulation study 2: Spatial Effects 50
3.2 Simulation study 2: Averaged fits of the nonparametric functions. 52
3.3 Food, kitchen supplies & water: Estimated nonparametric effects in the selection and outcome equation. 56
3.4 Construction material & tools: Estimated nonparametric effects. 57
3.5 Food, kitchen supplies & water: Estimated spatial effects 61
3.6 Construction material & tools: Estimated spatial effects. 62

4.1 Joint and marginal densities in one Monte Carlo draw of simulation setting (iii) and setting (iv). 82
4.2 Setting (b.iii): Estimated curves in first 50 simulation runs for $n = 100$. . 90
4.3 Estimated effects for 4th and 5th grade students. 92
4.4 Estimated marginal and joint error densities for 4th and 5th grade students. 93

List of Tables

2.1 Coverage rates in simulations together with average areas in parenthesis. . 25
2.2 Coverage rates and (average areas) for Simulation 1 with correlated covariates. 26
2.3 Parametric estimates. 31

3.1 Simulation study 1: Averaged estimation bias in the cases of correlated and identical design matrices. 46
3.2 Simulation study 1: Empirical root mean squared errors. 48
3.3 Simulation study 1: Estimation bias and root mean squared errors for the correlation between the errors and the variance in the outcome equation. . 48
3.4 Simulation study 2: Empirical root mean squared errors for univariate regressions and the sample selection model. 52
3.5 Parametric estimates. 55

4.1 Parametric simulation setting (i): Bivariate normality. 84
4.2 Parametric simulation setting (ii): Bivariate normality with outliers. . . 85
4.3 Parametric simulation setting (iii): Mixture of bivariate normals (unobserved clusters). 85
4.4 Parametric simulation setting (iv): Nonlinear conditional mean. 86
4.5 Setting (a) and (b): DGPs of Su & Ullah (2008) 88
4.6 Settings (b.ii), (b.iii) and (b.v): More complex distributions 89

1 Introduction

1.1 Motivation and Objectives

Regression techniques are among the principal tools of empirical scientists. Thereby, we aim at inferring from a set of independent variables (also called covariates or regressors) on a dependent measurement (the response variable). In applied research, most models simplify the relationships between dependent and independent variables to be parametric, i.e. it is *a priori* assumed that they are fully described by a finite set of parameters (of known dimension). However, such restrictive parametric models are rarely – or even *almost never* as Yatchew (1998) puts it in a well-cited article – justified by subject-matter theory, and can lead to seriously misleading inference if they are incorrect. Nonparametric regression techniques aim at relaxing these restrictions. In principle, these techniques neither make an assumption on the functional form (besides a smoothness condition) of an effect of a regressor nor the type and order of interactions between variables. It is well known that in high-dimensional settings impractically large data sets are then required due to the so-called curse of dimensionality. Therefore, the dimension has to be reduced and (semiparametric) additive models assuming additive separability of covariate effects (with possible two-dimensional interaction surfaces) have proven to be valuable in practice and are considered throughout this thesis. Specifically, we focus on variants of the (structured) additive model of the form

$$y_i = u_i'\gamma + f_1(x_{1i}) + \ldots + f_{p_1}(x_{p_1,i}) + x_{p_1+1,i}g_1(t) + \ldots + x_{p_1+p_2,i}g_{p_2}(t) + f_{\text{spat}}(s_i) + \varepsilon_i, i = 1, \ldots, n \tag{1.1}$$

where $u_i'\gamma$ corresponds to usual parametric effects and $f_1(\cdot),\ldots,f_{p_1}(\cdot)$ are smooth but otherwise unspecified functions of continuous covariates $x_{i1}, \ldots, x_{ip_1}$. Time-varying effects $g_1(t),\ldots,g_{p_2}(t)$ of covariates $x_{i,p_1+1},\ldots,x_{i,p_1+p_2}$ and a spatial effect $f_{\text{spat}}(s_i)$ of a regional variable s_i are only considered in Chapter 3 (although they could also be supported in the remaining chapters). ε_i is an unobserved error term commonly assumed to satisfy the conditional mean restriction $\text{E}(\varepsilon_i) = \text{E}(\varepsilon_i|u_i, x_{1i}, \ldots, x_{p_1+p_2,i}, s_i, t) = 0$. For inference but not necessarily for estimation often they are further assumed to be

independently and normally distributed with constant variance σ^2, i.e. $\varepsilon_i \sim \mathrm{N}(0, \sigma^2)$. These assumptions, however, are often not fulfilled in data situations in practice frequently limiting the use of standard nonparametric techniques and their software implementations. This may partly explain why the bulk of applied research still relies on parametric models where methods with weaker assumptions on the error term are more widely available. Further, although properties of nonparametric techniques are theoretically well understood, a lack of integrated powerful and reliable inferential tools in common implementations might frequently form an obstacle to the use of these methods. The aim of this thesis are thus to provide methods and implementations for estimation of additive models relaxing some assumptions usually imposed in available approaches and to provide tools for inference, i.e. means for quantification of estimation uncertainty and significance tests. Specifically, the objectives are

- the development of flexible methods for estimation and inference in various complex data situations. While all considered models allow to additively include smooth covariate effects, in the different chapters generalizations of common assumptions in model (1.1) are considered. In particular, in Chapters 3 and 4, flexible nonparametric Bayesian methods in the presence of nonrandom sampling and endogenous covariates are introduced, respectively.

- the provision of simultaneous confidence (credible) bands for all considered models. These bands allow to appropriately quantify the estimation uncertainty of function estimates. They can be used for assessing the statistical significance of an effect and for hypotheses on its functional form. Further, a novel nonparametric specification test is introduced in Chapter 2.

- the provision of easy accessible implementations of the models for computation in a broadly automated fashion. To this end, all proposed methods are implemented in easy-to-use R packages. Chapter 5 is devoted to their description.

- the investigation of the finite sample properties of the proposed approaches via Monte Carlo simulations.

- last but not least the study of questions ranging from needs-relatedness of relief supply in earthquake-affected communities accounting for temporal and spatial dynamics in Pakistan over determinants of childhood undernutrition in Kenya to the relationship between class sizes and scholastic achievements of students in Israel.

In the subsequent subsection, relevance of nonparametric estimation even in seemingly simple shapes such as U-shaped curves is demonstrated followed by a brief introduction to nonparametric regression. Then, an outline of the thesis closes the introduction.

1.2 A Case for Nonparametric Regression

Typical examples for relatively simple nonlinear relationships in (development) economics are diminishing returns and the ever-recurring (inverse) U-shaped hypothesis. In the latter, (economic) theory predicts some turning point in the relationship of a covariate and a response before and after which the response falls (rises) and then rises (falls) again, respectively. Interest of research is validation of the hypothesis and identification of the turning point. Most prominent examples of such hypotheses are inequality and environmental Kuznets curves. Specifically, the inequality Kuznets curve (Kuznets, 1955) postulates that inequality rises and then falls again with the increase of income per capita. Similarly, the environmental Kuznets curve (see e.g. Stern, 2004) suggests that indicators of environmental degradation first rise and then fall with increasing income. Since certainly such trends may be intuitively outlined by a quadratic curve, most studies proceed by approximating the relationship by a quadratic function of the explanatory variable (income) in a regression analysis. Then, conclusions are usually based on the statistical significance of the quadratic term and on the prediction of the turning point by the resulting regression coefficients. However, such a proceeding is potentially hazardous and misleading inference due to model misspecification can result. In fact, theory usually only predicts that the relationship will be smooth and monotonically increasing and decreasing before and after the turning point, respectively. Very rarely theory gives guidance on the shape of the curve, such as a linear first derivative and symmetry as given by a quadratic function.

Simulated data examples in Figure 1.1 illustrate possible pitfalls when approximating the relationship by a quadratic function. In Figure 1.1(a), we see that a quadratic function cannot unbiasedly capture a relationship which is monotone and smooth before and after the turning point but with an upward trend that gets stronger towards the turning point and thus does not have a linear first derivative. A quadratic model fitted to the simulated data resulted in an insignificant (p-value 0.776) quadratic term and suggested a U-shaped (instead of inverted U-shaped) relationship and thus did not predict the true turning point at all. Likewise, a skewed convex curve where growth is slower before the turning point than the decline afterwards (as in Figure 1.1(b)) cannot be properly predicted by a quadratic trend. Although a significant quadratic term was found, the skewness led to a predicted turning point that is considerably before the true one. Figures 1.1(c) and (d)

provide examples where quadratic models yield (significant) quadratic terms and predict turning points although there are none.
Note that quadratic relationships are also used to model diminishing returns (i.e. data situations as in Figure 1.1(c) and hypotheses without turning point). The same issue as described before applies here: A quadratic function restricts the first derivative of the apparently simple relationship to be linearly decreasing to zero (the turning point of the quadratic function). When the turning point lies within the data range, the model is misspecified afterwards. Further, in case of a nonconstant first derivative, estimated coefficients will be biased and inference (tests) invalid.
Of course, it is no news that the parametrization of the relationship influences the results (see for example Anand & Kanbur (1993) and Harbaugh, Levinson & Wilson (2002) for discussions with respect to inequality and environmental Kuznets curves, respectively). To deal with this, then usually specification searches over different parametric model specifications (e.g. polynomials) are carried out in order to avoid the specification error. Of course, specification searches are of great importance in many situations. However, they have several drawbacks. First of all, the number of specifications is usually quite restricted and thus the specification search might not include the right model. Secondly, the used model selection criterion might not choose the right one (or competing criteria might select different models). Finally, uncertainty due to model selection will be neglected in the finally chosen model which invalidates statistical theory.
In contrast, in nonparametric estimation, the relationship is allowed to be very flexible imposing only smoothness (ideally controlled by some data-driven criterion) in order to limit the variance of the estimate. Thus, a specification search with its drawbacks is mostly avoided. Nonparametric estimation was capable of properly capturing the relationships and predicting the turning points (when appropriate) in all of the simulated data examples.
Misspecified parametric models also affect the validity of significance tests. Let us consider the model $y = f(income) + \varepsilon$. At first we want to know in fact whether or not there is a significant deviation from a linear relationship between y and $income$. That is, we are interested in the null hypothesis $H_0 : \ f(income) = \gamma_1 income$ (i.e. that $f(\cdot)$ is a linear function) versus the alternative $H_1 : \ f(income) \neq \gamma_1 income$. However, in the procedure described above, we rather test against the alternative $H_1^\star : \ f(income) = \gamma_1 income + \gamma_2 income^2$, i.e. a parametric alternative which is only a single special case of H_1. Now, since inference treats the model as if it were exact, the test based on the parametric model cannot distinguish between a relationship with turning point and the important case of a slope converging to a horizontal line, for example. This discrimination is of particular interest in the analysis of the environmental Kuznets curve in the

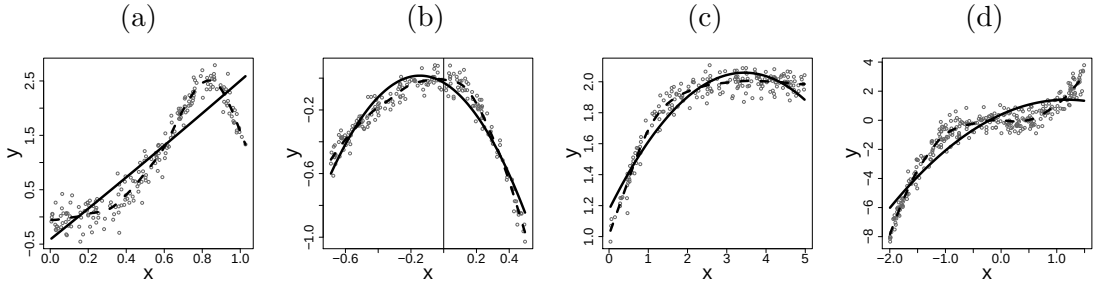

Figure 1.1: Examples of non-quadratic relationships in simulated data. In (a) and (b), relationships are smooth and monotone before and after a turning point. In the relationships in (b) and (c), no turning points are present. Dashed and solid lines indicate the true curves and predicted curves by quadratic models, respectively. In (b), the vertical line indicates the true turning point.

"race to the bottom" scenario (Dasgupta, Laplante, Wang & Wheeler, 2002). Therefore, one preferred strategy would be to actually test against the alternative H_1 using a nonparametric specification test and to then potentially inspect the derivatives of nonparametric estimates (and the corresponding simultaneous confidence bands in order to check the significance of a possible turning point). In Chapter 2, a powerful nonparametric specification test is proposed and applied to the investigation of U-shaped hypotheses on the impacts of the mother's nutritional status and the mother's height on child growth. More nonlinear relationships which can hardly be explained by parametric specifications are found throughout the thesis.

1.3 Nonparametric Regression in a Nutshell

The ultimate goal of nonparametric regression is to estimate the mean of a dependent variable y conditioned on covariates $x_{1i}, \ldots, x_{pi}$ in the model

$$y_i = f(x_{1i}, \ldots, x_{pi}) + \varepsilon_i, \ i = 1, \ldots, n$$

where $f(x_{1i}, \ldots, x_{pi})$ is a multidimensional unspecified function of interest describing the relationship between y_i and the covariates. ε_i is assumed to be an error term capturing dependencies between y_i and unknowns not included in (and uncorrelated with) $x_{1i}, \ldots, x_{pi}$. We first note that this implies that without further assumptions if $\mathrm{E}(\varepsilon_i | x_{1i}, \ldots, x_{pi}) \neq 0$ it follows that $\mathrm{E}(y_i | x_{1i}, \ldots, x_{pi}) \neq f(x_{1i}, \ldots, x_{pi})$ which is commonly known as the endogeneity problem in econometrics (and mostly confounding in

other disciplines) and will be of further interest later in this thesis. However, for the remainder of this introduction we assume that the equality holds.
Secondly, in practice, if p is large (usually already in case of $p > 2$) nonparametric estimation of $f(x_{1i}, \ldots, x_{pi})$ becomes intractable since the amount of data needed to obtain a desirable accuracy grows exponentially with p which is commonly referred to as the *curse of dimensionality*. Structured additive models like the one given in Equation 1.1 aim at mitigating this problem. These are still more flexible than parametric models but reduce the dimension of fully nonparametric models by making the additivity assumption $f(x_{1i}, \ldots, x_{pi}) = \sum_{j=1}^{p} f_j(x_{ji})$. Moreover, they facilitate graphical representation and interpretability of the results. See Hastie & Tibshirani (1990) and Wood (2006) for extensive treatments of additive models. Many techniques for estimation of such models exist and only some of them will be mentioned here.
A first class of nonparametric estimators are local smoothers of which *nearest neighbors*, *locally weighted regression (Loess)* and *local polynomial (kernel) regression* (with the Nadaraya-Watson estimator as the most well-known special case) are prominent members. The idea of *local polynomial regression* in the one-dimensional case is to approximate the curve $f(x)$ at some point x by locally fitting a polynomial of degree d in the neighborhood of x such that less weight is assigned to observations far from x. To do so, the weighted least squares criterion is minimized

$$\min_{\gamma_0,\ldots,\gamma_d} \sum_{i=1}^{n} \left\{ y_i - \sum_{l=0}^{d} \gamma_l (x_i - x)^l \right\}^2 K\left(\frac{x_i - x}{\lambda}\right) \tag{1.2}$$

with some kernel function $K(\cdot)$ (e.g. the standard Gaussian density) and bandwidth parameter λ controlling how quickly the weights tend to zero. In the multivariate framework $p > 1$, a backfitting-algorithm can be employed. The idea of backfitting is to obtain estimates for $f_1, \ldots, f_p$ by iteratively smoothing the partial residuals for one $f_j,\ j = 1, \ldots, p$ in each step until the individual functions don't change. That is, after initialization, to approximate f_j we replace y_i in (1.2) by $y_i - \gamma_0 - \sum_{j=1}^{p} \hat{f}_j(x_{ji})$ and cycle through $j = 1, \ldots, p$ until convergence.
The advantages of local polynomial smoothers include their well-known theoretical properties (see e.g. Fan & Gijbels (1996) for an overview). However, in the additive model framework, backfitting particularly complicates the construction of inferential tools for these models which has led to the development of more complicated approaches.
In contrast, spline based procedures largely allow the direct fitting of additive models by penalized least squares making them an attractive alternative in multidimensional frameworks. Thereby, instead of the local formulation of the regression problem, a

global optimization problem is formulated. For $p = 1$, the *smoothing spline estimator* is the minimizer of the penalized least-squares criterion

$$\min_{f \in C^q} \left[\sum_{i=1}^{n} \{y_i - f(x_i)\}^2 + \lambda \int \{f^{(q)}(x)\}^2 dx \right] \tag{1.3}$$

over all q-times continuously differentiable functions f. The first term is the residual sum of squares measuring closeness to the data and the second term penalizes curvature in the functions. The latter is controlled by smoothing parameters $\lambda > 0$ similar to the bandwidth in kernel regression with smoother fits for large values of λ and more wiggly curves for small values of λ. Commonly, $q = 2$ is of interest in which case it turns out that the unique solution to (1.3) is the natural cubic spline with knots equal to the observations. The latter implies that direct fitting of smoothing splines is computationally intensive in the multivariate setting because high dimensional matrices have to be inverted. *Penalized splines* (see e.g. Ruppert, Wand & Carroll, 2003) can be considered as a low-rank generalization of smoothing splines and enjoy increasing popularity in recent years last but not least due to their advantages in additive models (see Equation (2.2) for the optimization criterion). They allow to flexibly choose the number of knots (typically considerably smaller than n), the spline basis (e.g. B-splines or truncated polynomials) and the penalty (e.g. integrated squared derivatives of the spline functions or their approximation by difference penalties). We will focus exclusively on Bayesian and frequentist versions of penalized splines in this thesis and describe them in more detail in the following chapters.

1.4 Outline

In **Chapter 2**, we study the determinants of chronic undernutrition (measured by the WHO stunting Z-score, see WHO, 2006) of Kenyan children, with particular focus on the highly non-linear age pattern in undernutrition. In particular, we are interested in the possibility of catch-up growth, i.e. improvements of the nutritional status over age. This is complicated by the fact that the age curve exhibits considerable functional heterogeneity, i.e. the degree of smoothness of the curve varies over age with a rapid deterioration in the first year of life and a relatively constant curse afterwards. This cannot be captured by a usual global smoothing parameter (controlling the degree of penalization of the roughness of the curve), but has to be modeled by "locally adaptive smoothing". Further, we are interested in the shapes of the impacts of the mother's body mass index and her height, which were found to be inverse U-shaped in previous studies.

To answer these questions, simultaneous confidence bands for additive models with locally-adaptive smoothed components and heteroscedastic errors are proposed. These appropriately quantify the estimation uncertainty of function estimates and can be used for assessing the statistical significance of an effect and for hypotheses on its functional form. Further, a novel nonparametric specification test is introduced which is used for the latter question where we are interested in the relevance of a deviation from a linear specification of the effects. The confidence bands and the specification test are shown to perform very well in extensive Monte Carlo simulations.
We find a statistically significant improvement of the stunting score between ages of 23 and 28 months which, however, is shown to be most likely picking up the fact that children younger than 2 years were measured recumbent and children older than 2 years were measured standing. A possible pitfall in the construction of the stunting Z-score is revealed despite the extreme noisiness of the data which renders the comparison to the implied reference population of healthy children problematic. As a consequence of the construction of the stunting Z-score, the aggregated measure of stunting might underestimate the state of chronic undernutrition in the country. Our analysis emphasizes the importance of nonparametric estimation of the age effect in order to avoid misspecification bias in fully parametric models.
While in Chapter 2, the data is assumed to be randomly sampled, in **Chapter 3** we consider the case where observations are made non-randomly according to some selection mechanism described by an additional regression equation (explaining the selection probability). In the case of correlations between unobservable determinants of the selection probability and unobservables influencing the variable of primary interest, standard regression techniques yield biased estimates and (parametric) sample selection models are usually applied. We propose a flexible Bayesian approach to correct for the sample selection bias and model temporal and spatial dynamics of relief supply in earthquake affected regions in Pakistan. Thereby, the decision to deliver goods and the factors that determine the amount of goods supplied are analyzed simultaneously. Interesting results include that effects of needs-related variables show a strong time dependence suggesting organizational learning in the humanitarian community. Further, spatial patterns are recovered that go beyond what heterogeneity in local damage can explain.
In **Chapter 4**, we relax the usual assumption in Equation 1.1 that
$\mathrm{E}(\varepsilon|u, x_1, \ldots, x_{p_1}) = 0$ and allow one of the explanatory variables to be correlated with the unobservable error term relying on the availability of an instrumental variable. A violation of this assumption is prevalent particularly but not exclusively in the social sciences in the case of non-experimental data where the correlation between regressors and error term may result from confounders (omitted variables), measurement error,

reverse causality and sample selection, for example. It is well-known that standard regression techniques then yield biased estimates and instrumental variable regression to correct for endogeneity bias is commonly applied. We propose a Bayesian nonparametric instrumental variable approach where bias correction relies on a simultaneous equations specification with flexible modeling of both the covariate effects and the joint error distribution. This allows us to construct simultaneous credible bands (the Bayesian analogue to confidence bands) without distributional assumption on the error terms. The approach is used for the analysis of the relationship between class size and scholastic achievements of students in Israel.
Finally, **Chapter 5** is devoted to the practical use of the R packages providing implementations of all methods proposed in the thesis.
The thesis is based on the following papers:

- Wiesenfarth, M. and Kneib, T. (2010). Bayesian Geoadditive Sample Selection Models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59* (3), 381–404.

- Wiesenfarth, M., Krivobokova, T., Klasen, S. and Sperlich, S. (2012). Direct Simultaneous Inference in Additive Models and its Application to Model Undernutrition. *Journal of the American Statistical Association*, forthcoming.

- Wiesenfarth, M., Hisgen, C. M., Kneib, T. and Cadarso-Suarez, C. (2012). Bayesian Nonparametric Instrumental Variable Regression based on Penalized Splines and Dirichlet Process Mixtures. *Working Paper*.

2 Direct Simultaneous Inference in Additive Models and its Application to Model Undernutrition

Abstract: *This chapter proposes a simple and fast approach to build simultaneous confidence bands and perform specification tests for smooth curves in additive models. The method allows for handling of spatially heterogeneous functions and its derivatives as well as heteroscedasticity in the data. It is applied to study the determinants of chronic undernutrition of Kenyan children, with particular focus on the highly non-linear age pattern in undernutrition. Model estimation using the mixed model representation of penalized splines in combination with simultaneous probability calculations based on the volume-of-tube formula enable the simultaneous inference directly, i.e. without resampling methods. Finite sample properties of simultaneous confidence bands and specification tests are investigated in simulations. To facilitate and enhance its application, the method has been implemented in the R package AdaptFitOS.*

2.1 Introduction

In empirical studies one is typically interested not only in estimation of parameters or curves, but also in statistical inference about these estimators. Constructing confidence intervals and performing corresponding specification tests are necessary tools for going beyond the first steps of data exploration. Compared to the finite-dimensional parametric case, inference about a smooth function f, say, in the univariate nonparametric regression context is much more involved. The pointwise confidence bands for $f(x)$ that are usually given do not assess the whole function. Another commonly used confidence band based on Bayesian smoothing splines proposed by Wahba (1983) (see also Nychka, 1988) is only valid in the average coverage sense. That is, the nominal coverage probability results by averaging the coverage probabilities for $f(x)$ at each sample point, so

that the confidence band is valid neither at each point nor for the entire curve simultaneously. In general, both pointwise and Wahba (1983)'s confidence bands do not permit statements about the statistical significance of certain features in the underlying curve. Instead, one needs a simultaneous confidence band for f (from some suitable class of functions $\mathcal{F}$, say), which is typically based on its nonparametric estimator $\hat{f}$, and is given by $\left\{ \hat{f}(x) - c\sqrt{\mathrm{Var}\{\hat{f}(x)\}},\ \hat{f}(x) + c\sqrt{\mathrm{Var}\{\hat{f}(x)\}},\ \forall x \in \mathcal{X} \right\}$, where c satisfies

$$\alpha = \inf_{f \in \mathcal{F}} P_f \left(\frac{|\hat{f}(x) - f(x)|}{\sqrt{\mathrm{Var}\{\hat{f}(x)\}}} > c, \forall x \in \mathcal{X} \right)$$

on some subspace of the predictor space $\mathcal{X}$ for a given $\alpha \in (0, 1)$. Such a confidence band can be used, for example, in tests for functional form specification. Note that c depends crucially on f, which is unknown in practice.

There is an extensive theoretical literature on simultaneous confidence bands for models with a single curve. In a seminal paper, Bickel & Rosenblatt (1973) relate the asymptotic distribution of $\sup_{x \in \mathcal{X}} \left| \hat{f}(x) - \mathrm{E}\{\widehat{f}(x)\} \right|$ (that is, ignoring the bias $\mathrm{E}\{\widehat{f}(x)\} - f(x)$ that depends on the unknown f) to the distribution of the supremum of a Gaussian process. However, the convergence of these normal extremes is known to be exceedingly slow with $\log(n)^{-1}$ for sample size n, resulting in very poor performance in small samples. This has led to the development of confidence bands based on bootstrapping techniques in combination with slight undersmoothing, see for example Neumann & Polzehl (1998) and Claeskens & Van Keilegom (2003). In general, such resampling methods are extremely numerically demanding and the data-driven choice of an appropriate smoothing parameter is still an open (and difficult) issue. Hence, in applications with large number of observations and a complicated model structure bootstrapping techniques introduce an unacceptable computational burden.

For our study of undernutrition of children in Kenya we are confronted with a data set of nearly $5,000$ observations. The aim is to investigate the relationship between the so-called Z-score for height for age measuring chronic undernutrition (often called 'stunting') typically used by the WHO (see e.g. WHO, 1995) and various continuous covariates, modeled additively. Initial explorative analysis has indicated heteroscedasticity in the data and has shown that at least one component of the model needs to be estimated using locally adaptive methods. Such a task is hardly feasible for bootstrap based techniques.

Another approach to building simultaneous confidence bands is to consider the *tail probabilities* of suprema of Gaussian random processes, exploring its connection to the so-called volume-of-tube formula, see Sun (1993), Sun & Loader (1994) and Johansen &

Johnstone (1990). As long as f can be estimated without a bias, this method yields very good results for $c \to \infty$ even in small samples, making resampling methods redundant. Recently, Krivobokova, Kneib & Claeskens (2010) have shown that using the mixed model representation of penalized splines (for a comprehensive overview see Ruppert, Wand & Carroll, 2003) for the curve estimation in combination with the approach of Sun (1993) has several advantages compared to other available techniques. However, they only consider univariate models with homoscedastic errors and do not allow for functional heterogeneity. Certainly, in practice usually more complicated data situations arise which limits the use of their approach. Motivated by such a complex data set concerning stunting by age in Kenya, our work aims at filling this gap. Specifically, we extend the approach of Krivobokova, Kneib & Claeskens (2010) to much more involved additive models with heterogenous functional components and heteroscedastic errors. Further, a completely new specification test for the components of an additive model is introduced that naturally takes a possibly varying residual variance as well as spatial heterogeneity of additive model components into account.

Simultaneous inference in additive models has to date not received much attention in the literature. Härdle, Huet, Mammen & Sperlich (2004) developed simultaneous confidence bands and specification tests for generalized additive models in the kernel regression context. Wang & Yang (2009) propose an oracally efficient spline-backfitted kernel smoothing estimator for additive models and obtain asymptotic simultaneous confidence bands around the additive components using results for kernel regression in line with Bickel & Rosenblatt (1973). The main contribution of this work is an efficient estimation procedure with preliminary spline smoothing followed by univariate kernel regression, which allows for fast calculations. Extensions to additive autoregression models are pursued in Wang & Yang (2007) and in Song & Yang (2010), while Ma & Yang (2011) treated partially linear additive models. Härdle, Sperlich & Spokoiny (2001) proposed locally adaptive (via wavelets) and bandwidth adaptive specification tests for additive models. In our work, we employ penalized splines for estimation which avoids backfitting or marginal integration in additive models and allows to obtain (adaptive) smoothing parameters from the corresponding (restricted) likelihood simultaneously with the main parameters of interest. Moreover, estimation of the varying residual variance can be incorporated with little additional numerical effort. The main advantage of the method we propose in this chapter is that one can obtain simultaneous confidence bands with very good small sample properties for sophisticated models – such as additive models with heterogeneous smooth components and heteroscedastic errors – instantly, i.e. without resampling methods. Simple and fast calculations allow us also to perform model selection and specification tests in seconds. The approach is implemented in the R package

AdaptFitOS, making it readily available for practitioners.
The chapter is organized as follows. In Sections 2.2 and 2.3 additive models with penalized splines and the data are introduced. In Section 2.4 uniform confidence bands are considered, while a new model specification test is proposed in Section 2.5. The performance of our approach is investigated in Monte Carlo simulations in Section 2.6. The methods are used then to analyze the determinants of undernutrition of children in Kenya in Section 2.7 before we conclude in Section 2.8. Some of the technical details are deferred to the Appendix.

2.2 Additive Models with Penalized Splines

Let us start with a simple additive model

$$Y_i = \beta_0 + \sum_{j=1}^{d} f_j(x_{ji}) + \varepsilon_i, \quad \varepsilon_i \sim \mathcal{N}(0, \sigma^2), \;\; i = 1, \ldots, n, \tag{2.1}$$

where the constant β_0 is an intercept. Without loss of generality we assume non-random covariates to be scaled to the unit interval, i.e. $x_{j1}, \ldots, x_{jn} \in [0,1]$, $j = 1, \ldots, d$. Each corresponding $f_j \in C^q[0,1]$ is a q times continuously differentiable function and is centered at zero to ensure identifiability, i.e. we assume $\mathrm{E}\{f_j(x_j)\} = 0$. To estimate f_j with penalized splines, we define for each f_j, $j = 1, \ldots, d$ a set of $k_j < n$ knots $\tau_j = \{0 < \tau_{j,1} < \ldots < \tau_{j,k_j} < 1\}$ and denote the corresponding spline space of degree p as $\mathcal{S}(p; \tau_j)$. This set consists of $p-1$ times continuously differentiable functions, that are polynomials of degree p on each $[\tau_{j,i}, \tau_{j,i+1})$. Then, the penalized spline estimator is the solution to

$$\min_{s_j(x) \in \mathcal{S}(p;\tau_j),\, j=1,\ldots,d} \left[\sum_{i=1}^{n} \left\{ Y_i - \beta_0 - \sum_{j=1}^{d} s_j(x_{ji}) \right\}^2 + \sum_{j=1}^{d} \lambda_j \int_0^1 \{s_j^{(q)}(x)\}^2 dx \right], \tag{2.2}$$

for some $q \leq p$. Claeskens, Krivobokova & Opsomer (2009) studied asymptotic properties of univariate penalized spline estimators under very mild regularity conditions on the distribution of the covariates and knots, which are further assumed to hold for (2.2) as well. Note also that all subsequent results are directly adjustable to random designs. In principle, one can choose different spline degrees for each $\mathcal{S}(p; \tau_j)$ and different penalization orders q for each s_j, but we do not consider this generalization here. To solve (2.2), represent each $s_j(x)$ as a linear combination of $k_j + p + 1$ spline functions that form basis in $\mathcal{S}(p; \tau_j)$. We use B-splines in our implementation, although others are

also certainly possible. Denote a row vector $B_j(x) = \{B_{j,1}(x, \tau_j), \ldots, B_{j,k_j+p+1}(x, \tau_j)\}$ to be some spline basis for $\mathcal{S}(p; \tau_j)$ and let $B_j = \{B_j(x_{j1})^t, \ldots, B_j(x_{jn})^t\}^t$ be the corresponding basis matrix. To obtain centered estimates for f_j, one uses the centered basis matrix $\tilde{B}_j = (\mathrm{I}_n - 1_n 1_n^t)B_j$, with 1_n as an n-dimensional column vector of ones. Now, representing each $s_j(x) = \tilde{B}_j(x)\beta_j$ allows to solve (2.2) as a minimization problem over β_j.

Smoothing parameters λ_j can be chosen using multivariate versions of cross-validation. An alternative way to estimate smoothing parameters λ_j is to exploit the link between penalized splines and linear mixed models. Decompose each $\tilde{B}_j\beta_j = \tilde{B}_j(F_b^j b_j + F_u^j u_j) = X_j b_j + Z_j u_j$ in such a way that $(F_u^j)^t F_b^j = (F_b^j)^t D_j F_b^j = 0$ and $(F_u^j)^T D_j F_u^j = I_{\tilde{k}_j}$, where D_j is such that $\int_0^1 [\{\tilde{B}_j(x)\beta_j\}^{(q)}]^2 dx = \beta_j^t D_j \beta_j$ and $\tilde{k}_j = k_j + p + 1 - q$. This decomposition is not unique due to singularity of D_j. In our implementation we followed Durban & Currie (2003). Assuming

$$Y|u_1, \ldots, u_d = \beta_0 + \sum_{j=1}^{d}(X_j b_j + Z_j u_j) + \varepsilon, \;\; u_j \sim \mathcal{N}(0, \sigma_{u_j}^2 I_{\tilde{k}_j}), \;\; j = 1, \ldots, d, \qquad (2.3)$$

for $Y = (Y_1, \ldots, Y_n)^t$ and $\varepsilon \sim \mathcal{N}(0, \sigma^2 I_n)$ leads to the standard linear mixed model with the best linear unbiased predictor being equal to the solution of (2.2) with $\lambda_j = \sigma^2/\sigma_{u_j}^2$. All mixed model parameters, including $\sigma^2/\sigma_{u_j}^2$, are estimated simultaneously by maximizing a single (restricted) likelihood function. In our further developments we will use the estimators for f_j that result from the mixed model representation of penalized splines (2.3), so that our estimator will have the form $\hat{f}_j(x) = \ell_j^t(x)Y$, with the smoothing matrix $\ell_j(x)$ given by

$$\ell_j(x) = (I - S_{-j})C_j\{C_j^t(I - S_{-j})C_j + \Lambda_j\}^{-1}C_j^t(x), \qquad (2.4)$$

where model matrix $C_j = [X_j \; Z_j]$, penalty matrix $\Lambda_j = \sigma^2/\sigma_{u_j}^2 \mathrm{diag}(0_q, 1_{\tilde{k}_j})$ and $S_{-j} = C_{-j}(C_{-j}^t C_{-j} + \Lambda_{-j})^{-1}C_{-j}^t$ with $C_{-j} = [C_1, C_2, \ldots, C_{j-1}, C_{j+1}, \ldots, C_d]$ and $\Lambda_{-j} = \mathrm{blockdiag}(\Lambda_1, \Lambda_2, \ldots, \Lambda_{j-1}, \Lambda_{j+1}, \ldots, \Lambda_d)$. For practical implementation standard mixed models software can be used (e.g. function `lme` in R).

2.3 Data on Childhood Undernutrition in Kenya

Using the model introduced in the previous section we aim to investigate the data on undernutrition of Kenyan children. Acute and chronic undernutrition is among the most serious health issues facing developing countries. It is not only an intrinsic indicator of

well-being but also associated with morbidity, mortality, reduced labor productivity, etc. Moreover, some estimates claim that undernutrition is implicated in more than 50% of deaths in developing countries (Pelletier, 1994). Given the importance of nutrition for child development, a particular focus is on promoting adequate nutrition for children. Consequently, there is an abundant theoretical and empirical literature on the determinants of childhood undernutrition in developing countries (see Horton, Alderman & Rivera, 2009). However, most studies are limited to parametric approaches or simple descriptive methods, not accounting for the complex functional forms of the relationships and neglecting the high uncertainty due to the large variability in the data (e.g. Kabubo-Mariara, Ndenge & Mwabu, 2009 and Victora, de Onis, Hallal, Blossner & Shrimpton, 2010).

We analyze the determinants of child undernutrition in Kenya, using the 2003 round of the Kenyan Demographic and Health Survey (KDHS2003, see Central Bureau of Statistics (CBS) Kenya, Ministry of Health (MOH) Kenya & ORC Macro, 2004). This includes information on $n = 4,561$ children, aged 0–60 months. The data are cross-sectional, i.e. there are no repeated observations of the same individual. We focus on the Z-score for stunting defined as

$$Z_i = \frac{H_i - \text{med}(H)}{\sqrt{\text{Var}(H)}},$$

where H_i is the height of the ith individual at a certain age and $\text{med}(H)$ and $\text{Var}(H)$ are the median and variance of the heights in a reference population of well-nourished and healthy children of the same age, respectively. By this normalization, a suitable Gaussian response is obtained and international comparability is aimed for. Note that our analysis is based on the new WHO child growth reference standard which was recently developed based on the assessment of child growth in healthy populations in six countries across the world. Roughly, as described in WHO (2006), to obtain $\text{med}(H)$ and $\sqrt{\text{Var}(H)}$ a generalized additive model for location, scale and shape (GAMLSS) was applied. Thereby, median heights and standard deviation were estimated as smooth functions of age using cubic splines with degrees of freedom chosen by (G)AIC. Since children younger than 2 years were measured recumbent and children older than 2 years were measured standing, 0.7 cm were added to all observations of children older than 2 years prior to fitting the model. This estimated difference of 0.7 cm was obtained as the mean differences between measurements of recumbent length and standing height of children between 18 and 30 months from which both measurements are available. Further, some power transformation was applied to age prior to fitting in order to expand the age scale for low age values and compress it for larger age values. This was necessary in

order to avoid oversmoothing for low age values where growth is much more rapid than for larger age values. After fitting, 0.7 cm were subtracted from the estimated median curve for all age values larger than 24 months.
Based on the literature on the determinants of chronic undernutrition (e.g. UNICEF, 1998), we start with the following simplified semiparametric model assuming i.i.d. Gaussian errors

$$Z_i = \beta_0 + f_1(age_i) + f_2(bmi_i) + f_3(mheight_i) + z_i'\gamma + \varepsilon_i,\ \varepsilon_i \sim N(0, \sigma^2),\ i = 1, \ldots, n, \quad (2.5)$$

where $f_1(age)$, $f_2(bmi)$ and $f_3(mheight)$ are smooth functions of the age of the child in months, the Body Mass Index (BMI, defined as weight in kg divided by the squared height in meters) of the mother and the mother's height, respectively. Constant smoothing parameters λ_j are assumed for all functions. Further, as control variables we add a set of covariates z including the numbers of years of education of the mother, the sex of the child as well as the location (rural/urban) and province of the household.
Some of the substantive questions for which a semi-parametric regression approach is particularly suitable concern the age effect. As shown in the literature on undernutrition (e.g. Belitz, Hübner, Klasen & Lang, 2010 and references therein), children in developing countries are usually born with an anthropometric status that is close to the median of the reference population. Due to poorer nutrition and a poorer health and sanitary environment, many children begin to fall behind, first in weight, and then in growth so that a growth deficit begins to emerge. This is usually intensified in the so-called weaning crisis, which ranges from 4 to 8 months of age, when solid foods and liquids are introduced and the poor quality of these foods and liquids in many poor countries worsens the nutritional status of the child. As children's bodies then partly adapt to poorer nutritional and health environment (largely by becoming more resistant to pathogens, partly by the reduced energy needs for a smaller body, and partly through lower activity levels), stunting usually stabilizes at around age 2, i.e. no further deteriorations vis-a-vis a reference population of healthy children is observed. One of the important questions in the literature concerns the possibility of catch-up growth (see e.g. WHO, 1995), i.e. improvements of the stunting Z-score over time, particularly after age 2. Thus an important empirical question to ask is in which countries and under which contexts such catch-up growth (usually assumed to be possible particularly between age 2 and 3) is observed. This amounts to testing whether the slope of the age effect is significantly above 0 in some interval.
A second substantive question concerns the impact of the mother's nutritional status, typically proxied by her BMI, on child growth. Some studies (see e.g. Kandala, Fahrmeir,

Klasen & Priebe, 2009) have found an inverse U-shape, where initially the BMI serves to improve the Z-score, but high levels of the BMI could signify poor quality nutrition which then leads to a worse nutritional status for the child. Again the shape of the curve is thus of interest here. Similar arguments can be made for the impact of mother's height on child height which is likely to be related to genetic transmission as well as inter-generational transmission of the economic status. Also here the shape of the curve is hard to guess in advance.
To answer these questions certain specification tests based on simultaneous confidence bands for additive models developed in the subsequent sections need to be employed.

2.4 Simultaneous Confidence Bands

2.4.1 The Volume-of-tube Formula

Sun & Loader (1994) suggested to build simultaneous confidence bands for a smooth function using the approximation to the *tail probability* of maxima of Gaussian random processes, which turned out to be connected to the volume-of-tube formula. In this case no bootstrap is necessary and the approach yields quite good results in small samples, once a function estimator is unbiased. For completeness we give here some details.
Consider model (2.1) with $d = 1$. Let $\tilde{f}(x) = \tilde{\ell}(x)^t Y$ be an unbiased estimator of f and assume λ to be known. This implies that $G(x) = \mathrm{Var}\{\tilde{f}(x)\}^{-1/2}\{\tilde{f}(x) - f(x)\} = \tilde{\ell}(x)^t\epsilon/\|\tilde{\ell}(x)\|$ is a zero mean Gaussian process with variance one and

$$\mathrm{Cov}\{G(x_1), G(x_2)\} = \left(\frac{\tilde{\ell}(x_1)}{\|\tilde{\ell}(x_1)\|}\right)^t \left(\frac{\tilde{\ell}(x_2)}{\|\tilde{\ell}(x_2)\|}\right) =: \eta^t(x_1)\eta(x_2),$$

with manifold $\{\eta(x) : x \in [0,1],\ \eta(x) = (\eta_1(x), \ldots, \eta_n(x))\}$. Then, according to Sun & Loader (1994), it holds for $c \to \infty$

$$\alpha = P\left(\sup_{x\in[0,1]} |G(x)| \geq c\right) = \frac{\kappa_0}{\pi}\exp\left(-c^2/2\right) + 2\{1 - \Phi(c)\} + o\left\{\exp(-c^2/2)\right\}, \quad (2.6)$$

with $\kappa_0 = \int_0^1 \|\frac{d}{dx}\eta(x)\|dx$ the length of the manifold $\eta(x)$ and $\Phi(\cdot)$ the distribution function of a standard normal distribution. With this, the $100(1-\alpha)\%$ simultaneous confidence band for $f(x),\ x \in [0,1]$ has the form

$$f(x) \in \left[\tilde{f}(x) - c\sqrt{\mathrm{Var}\{\tilde{f}(x)\}},\ \tilde{f}(x) + c\sqrt{\mathrm{Var}\{\tilde{f}(x)\}}\right], \forall x \in [0,1],$$

where c is found by inverting (2.6). In practice, however, all nonparametric estimators of f are biased and the smoothing parameter λ is estimated from the data, introducing extra variability. Both problems have been discussed in Krivobokova, Kneib & Claeskens (2010) who suggested (in the univariate case) to use instead of c a critical value c_m obtained from the mixed model representation of penalized splines (2.3). Some heuristic arguments and an extensive simulation study confirmed that this approach has very good small sample properties. Also, they showed that the variability due to the estimation of the smoothing parameter σ^2/σ_u^2 is negligible, once a small q is used. Thereby, one has to use enough knots (k proportional to $n^{\nu/(2q+1)}$, $\nu > 1$) to ensure that the approximation bias of the penalized spline estimator is negligible (approximation bias arises due to the fact that a smooth function f is replaced by a spline; it converges to zero with k^{-p-1}). For more details on the bias structure of penalized splines see Claeskens, Krivobokova & Opsomer (2009). In the next section we discuss how c_m can be obtained for a general additive model.

2.4.2 Simultaneous Confidence Bands for Additive Models

We consider model (2.3) and assume that sufficiently many knots are taken, so that the approximation bias is small enough and one can replace $f_j(x)$ by $X_j(x)b_j + Z_j(x)u_j =: C_j(x)\theta_j$ directly. To obtain $c_{m,j}$ we consider the marginal distribution of Y, that is

$$Y \sim \mathcal{N}\left(\beta_0 + \sum_{j=1}^{d} X_j b_j,\ \sigma^2 I_n + \sum_{j=1}^{d} \sigma_{u_j}^2 Z_j Z_j^t\right).$$

With respect to this distribution we obtain a zero mean Gaussian process

$$G_{m,j}(x) = \frac{C_j(x)(\hat{\theta}_j - \theta_j)}{\sqrt{C_j(x)\mathrm{Cov}(\hat{\theta}_j - \theta_j)C_j(x)^t}} \sim \mathcal{N}(0,1),$$

where $\mathrm{Cov}(\hat{\theta}_j - \theta_j) = \{C_j^t(I_n - S_{-j})C_j + \Lambda_j\}^{-1}$ and

$$\mathrm{Cov}\{G_{m,j}(x_1), G_{m,j}(x_2)\} = \left(\frac{\ell_{m,j}(x_1)}{\|\ell_{m,j}(x_1)\|}\right)^t \left(\frac{\ell_{m,j}(x_2)}{\|\ell_{m,j}(x_2)\|}\right) =: \eta_{m,j}^t(x_1)\eta_{m,j}(x_2),$$

with $\ell_{m,j}(x) = \{C_j^t(I - S_{-j})C_j + \Lambda_j\}^{-1/2} C_j^t(x)$. Since $G_{m,j}(x)$ is a zero mean Gaussian process, we can apply the volume-of-tube formula to obtain $c_{m,j}$ from

$$P\left(\sup_{x\in[0,1]} |G_{m,j}(x)| \geq c_{m,j}\right) = \frac{\kappa_{m,j}}{\pi} \exp\left(-c_{m,j}^2/2\right) + 2\{1 - \Phi(c_{m,j})\} + o\left\{\exp(-c_{m,j}^2/2)\right\}, \quad (2.7)$$

with $\kappa_{m,j} = \int_0^1 \| \frac{d}{dx}\eta_{m,j}(x)\| dx$ as the length of the mixed model manifold. Now a confidence band around f_j based on a penalized spline estimator $\hat{f}_j$ is built as

$$\left[\hat{f}_j(x) - c_{m,j}\sqrt{\mathrm{Var}\{\hat{f}_j(x)\}},\ \hat{f}_j(x) + c_{m,j}\sqrt{\mathrm{Var}\{\hat{f}_j(x)\}}\right],$$

where $\mathrm{Var}\{\hat{f}_j(x)\} = \sigma^2\|\ell_j(x)\|^2$ with $\ell_j(x)$ defined in (2.4). A careful check shows that the proofs in Krivobokova, Kneib & Claeskens (2010) carry over to our complex case. Hence, this confidence band should have coverage probability close to the nominal level without further corrections. The critical value $c_{m,j}$ is obtained directly from (2.7) and no bootstrap is necessary. The small sample performance of this band is investigated in Section 2.6.1.

2.4.3 Simultaneous Bands for Additive Models with Spatially Heterogeneous Components and Heteroscedastic Errors

So far we assumed a constant error variance σ^2. This assumption of homoscedasticity may often be violated when the variance changes with some covariate or depends on $E(Y)$. Further, we assumed constant smoothing parameters λ_j, which may be too restrictive for functions that exhibit strong spatial heterogeneity. For example, the mean function can change rapidly for low covariate values and remains rather constant afterwards. This makes it necessary to penalize little in one part of the covariate support and more severely in another, which is referred to as locally adaptive smoothing. To relax these assumptions, we define $u_{js} \sim \mathcal{N}\{0, \sigma_{u_j}^2(\tau_{j,s})\},\ s = 1, \ldots, k_j$ and $\varepsilon_i \sim \mathcal{N}\{0, \sigma^2(\tilde{x}_i)\},\ i = 1, \ldots, n$, where $\tilde{x}$ is one of the covariates or some linear combination of them. Assuming that the variance processes $\sigma_{u_j}^2(\tau_j)$ and $\sigma^2(\tilde{x})$ are smooth functions, we model them with penalized splines and estimate using the link to mixed

models. More precisely, we define a hierarchical mixed model

$$\begin{aligned} Y &= \beta_0 + \sum_{j=1}^{d}(X_j b_j + Z_j u_j) + \varepsilon, \quad \varepsilon|v \sim N(0,\ \sigma^2\Sigma_\varepsilon), \quad u_j|w_j \sim N(0,\ \Sigma_{u_j}), \\ \Sigma_\varepsilon &= \mathrm{diag}\{\exp(X_v\gamma + Z_v v)\}, \quad v \sim N(0,\ \sigma_v^2 I_{k_v}), \\ \Sigma_{u_j} &= \mathrm{diag}\{\exp(X_{w_j}\delta_j + Z_{w_j} w_j)\}, \quad w_j \sim N(0,\ \sigma_{w_j}^2 I_{k_{w_j}}), \end{aligned} \tag{2.8}$$

where X_v and Z_v are obtained by decomposing B_j in the same fashion as described in Section 2.2, but based on a smaller number of knots $k_v \ll k_j$. In contrast, X_{w_j} and Z_{w_j} are obtained by decomposing the basis matrix $B_j = \{B_j(\tau_{j,1}, \tau_{w_j})^t, \ldots, B_j(\tau_{j,k_j}, \tau_{w_j})^t\}^t$. This basis matrix is obtained by treating knots τ_j as observations and choosing as knots τ_{w_j} a smaller subset of τ_j. All parameters of this model can be estimated from the corresponding (restricted) likelihood. A similar idea was suggested in a fully Bayesian framework with $d = 1$ and MCMC techniques by Crainiceanu, Ruppert, Carroll, Joshi & Goodner (2007). To overcome the numerically intensive computations of the latter, Krivobokova, Crainiceanu & Kauermann (2008) suggested to use the Laplace approximation of the likelihood. They have shown, that the resulting estimator is nearly identical to the Bayesian one, but can be obtained with considerably smaller numerical effort.

In the following, we extend the method of Krivobokova, Crainiceanu & Kauermann (2008) to the model with heteroscedastic errors and provide some details on the estimation procedure. To keep the exposition as clear as possible we will cover the single covariate case with varying residual variance and constant smoothing parameter. Details on the estimation of a model with varying smoothing parameter and constant error variance are given in Krivobokova, Crainiceanu & Kauermann (2008). The combination of varying smoothing parameter with varying residual variance, as well as the extension to additive models, is straightforward. Thus, we provide details only on the estimation of the model

$$\begin{aligned} Y|u, v &\sim N(Xb + Zu, \sigma^2\Sigma_\varepsilon),\ u \sim N(0, \sigma_u^2 I_k), \\ \Sigma_\varepsilon &= \mathrm{diag}\{\exp(X_v\gamma + Z_v v)\},\ v \sim N(0, \sigma_v^2 I_{k_v}) \end{aligned} \tag{2.9}$$

The marginal likelihood of model (2.9) is given by

$$L(b, \gamma, \sigma_u^2, \sigma^2, \sigma_v^2) = (2\pi)^{-\frac{(n+k_v)}{2}} \sigma_v^{-k_v} \int_{\mathbb{R}^{k_v}} \exp\{-g(v)\} dv \tag{2.10}$$

where $2g(v) = \log|V| + v^t v/\sigma_v^2 + (Y - Xb)^t V^{-1}(Y - Xb)$, with $V = \sigma^2\Sigma_\varepsilon + \sigma_u^2 ZZ^t$.

Since the integral in (2.10) is not available analytically, we opt to use the Laplace approximation. This is justified because the approximation error is of order k_v/n (see Severini, 2000) and we assumed $k_v \ll n$. After applying the Laplace approximation, the log-likelihood corresponding to (2.9) results in

$$-2l(b,\gamma,\sigma^2,\sigma_v^2,\sigma_u^2) \approx k_v \log\sigma_v^2 + \log|V(\hat{v})| + \log|I_{vv}(\hat{v})| + \frac{\hat{v}^t\hat{v}}{\sigma_v^2} + (Y-Xb)^t V^{-1}(\hat{v})(Y-Xb),$$

with $\hat{v}$ as a solution to $0 = \partial g(v)/\partial v = -Z_v^t\left[\{(ZZ^t)^{-1}Z\hat{b}\}^2\sigma^2\Sigma_\varepsilon - \mathrm{diag}(A_1)\right]/2 + v\sigma_v^{-2}$, where A_1 denotes the vector of diagonal elements of matrix $\mathcal{A} = Z(Z^t\sigma^{-2}\Sigma_\varepsilon^{-1}Z + \sigma_u^{-2}I_k)^{-1}Z^t(ZZ^t)^{-1}\sigma_u^{-2}$. The corresponding Fisher information matrix is given by $I_{vv}(v) = \mathrm{E}\left(\partial^2 g(v)/\partial v\partial v^t|v\right) = Z_v^t\mathrm{diag}(A_2)Z_v/2 + \sigma_v^{-2}I_{k_v}$. Here, A_2 is the vector of diagonal elements of matrix $\mathcal{A}^2$. Introducing notations $\omega = (\gamma, v)$, $C_v = [X_v \;\; Z_v]$ and $D_v = \mathrm{diag}(0, I_{k_v})$, one can obtain estimates $\hat{\gamma}$ and $\hat{v}$ simultaneously from the iterated weighted least squares

$$\hat{\omega} = \frac{1}{2}\left(\frac{1}{2}C_v^t\mathrm{diag}(A_2)C_v + \sigma_v^{-2}D_v\right)^{-1} C_v\mathrm{diag}(A_2)\alpha, \tag{2.11}$$

with the working vector $\alpha = C_v\omega + \mathrm{diag}(A_2^{-1})\{(ZZ^t)^{-1}Z\hat{b}\}^2\sigma^2\Sigma_\varepsilon - \mathrm{diag}(A_1)$. The corresponding variance is estimated as

$$\hat{\sigma}_v^2 = \hat{v}^t\hat{v}/\mathrm{tr}\{Z_v^t\mathrm{diag}(A_2)Z_vI_{vv}^{-1}\}. \tag{2.12}$$

Thus, the parameters of model (2.9) can be estimated by iterating between estimation of $\hat{b}$, $\hat{u}$, $\hat{\sigma}^2$, $\hat{\sigma}_u^2$ for a fixed ω and σ_v^2 using standard linear mixed model software and updating $\hat{\omega}$ and $\hat{\sigma}_v^2$ from (2.11) and (2.12). To use the restricted likelihood, one has to replace $g(v)$ by $g_r(v) = g(v) + \log|X^tV^{-1}X|/2$.

The smoothing matrix for the penalized spline estimators in model 2.9 has now the form

$$\ell_j(x) = \Sigma_\varepsilon^{-1}(I - S_{-j})C_j\{C_j^t\Sigma_\varepsilon^{-1}(I - S_{-j})C_j + \Lambda_j\}^{-1}C_j^t(x) \tag{2.13}$$

with $\Lambda_j = \sigma^2\,\mathrm{blockdiag}(0_q, \Sigma_{u_j}^{-1})$ and $S_{-j} = C_{-j}(C_{-j}^t\Sigma_\varepsilon^{-1}C_{-j} + \Lambda_{-j})^{-1}C_{-j}^t\Sigma_\varepsilon^{-1}$. Note that $\mathrm{Var}\{\hat{f}_j(x)\} = \sigma^2\ell_j(x)^t\Sigma_\varepsilon\ell_j(x)$. Then, simultaneous confidence bands can be obtained as described in Section 2.4.2. Since k_{w_j} and k_v are both typically very small (5 – 10 subknots are usually sufficient), following the arguments of Krivobokova, Kneib & Claeskens (2010) one can show that the variability due to estimation of Σ_ε and Σ_{u_j} is negligible for sufficiently large n and small q. Our simulation study in Section 2.6.2 con-

firms this. The approach can also be used for investigating the statistical significance of features like dips and bumps. In order to do so, choose $q \geq 2$ and build the simultaneous confidence band around the estimated first derivative of f_j using

$$\ell'_j(x) = \Sigma_\varepsilon^{-1}(I - S_{-j})C_j\{C_j^t\Sigma_\varepsilon^{-1}(I - S_{-j})C_j + \Lambda_j\}^{-1}C_j'^t(x),$$

where $C_j(x)$ in (2.13) is replaced by the first derivative of the basis matrix $C'_j(x)$ (see Ruppert, Wand & Carroll, 2003, Chapter 6.8).Analogously, the critical value is obtained by replacing $\ell_{m,j}(x)$ by $\ell'_{m,j}(x)$.
Thus, using the mixed model representation of penalized splines one can estimate complex additive models with varying smoothing parameters and varying residual variance easily and obtain simultaneous confidence bands for the corresponding functions without additional effort.

2.5 A new Specification Test

The constructed simultaneous confidence bands can now be used for testing a parametric regression specification versus a quite general nonparametric alternative modeled by penalized splines. That is, we test the hypotheses

$$H_0: \ f_j(x) = f_j^0(x) \text{ vs } H_1: \ f_j(x) = f_j^0(x) + g_j(x), \ \forall x \in [0,1],$$

with $f_j^0(x)$ as a pre-specified polynomial function, whereas $g_j(x)$ is an unspecified deviation. The idea is to write $f_j(x) = f_j^0(x) + Z_j(x)u_j$ and to exploit the orthogonality of $f_j^0(x)$ and $Z_j(x)u_j$. Then, the above test is equivalent to testing $H_0: \ Z_j(x)u_j = 0$. This hypothesis can be checked by constructing a simultaneous confidence band around $g_j(x) = Z_j(x)u_j$. Since any spline function of degree q can be decomposed into a $q-1$ degree polynomial and a remainder, we can always choose such ψ_l that $f_j^0(x) = \sum_{l=1}^{q-1}\psi_l x^l$. Obviously, the test procedure corresponds to checking whether the confidence band for $Z_j(x)u_j$ uniformly encloses the zero line coinciding with the test statistic

$$T_j = \sup_{x\in[0,1]}\left(|Z_j(x)\hat{u}_j|/\sqrt{\text{Var}\{Z_j(x)\hat{u}_j\}}\right).$$

Rejection of H_0 takes place if $T_j > c^*_{m,j}$. The critical value $c^*_{m,j}$ and $\text{Var}\{Z_j(x)\hat{u}_j\} = \sigma^2\|\ell_j(x)\|^2$ are obtained by replacing C_j and C_{-j} in definitions (2.4) and (2.7) by $C_j := Z_j$ and $C_{-j} := [X_1, Z_1, \ldots, X_{j-1}, Z_{j-1}, X_j, X_{j+1}, Z_{j+1}, \ldots, X_d, Z_d]$, as well as appropriately adjusting Λ_j and Λ_{-j}. Adjustments to the cases of heteroscedastic errors

and locally adaptive smoothed components follow from the definitions in Section 2.4.3. Note that approximative p-values can be obtained by calculating the tail probabilities using the volume-of-tube formula (2.7) replacing $c_{m,j}$ by T_j.
By exploiting the decomposition of a spline function, improved power is obtained compared to the test strategy proposed in Claeskens & Van Keilegom (2003), for example. They build their proposed test on the simultaneous confidence band around f_j itself with the hypotheses $H_0: \ f_j(x) = f_j^0(x)$ vs $H_1: \ f_j(x) \neq f_j^0(x), \ \forall x \in [0,1]$, and rely on local polynomials for estimation and bootstrapping to obtain the critical value. Thereby, the data-driven choice of smoothing parameters is still an open problem.
Similar to our findings for the confidence bands, our test also has the advantage of performing well in small samples and of being analytically available, i.e. no bootstrap or Monte Carlo simulation is necessary (as in Härdle, Huet, Mammen & Sperlich, 2004, for example).In particular, this test is preferable to F-type tests as used in the R package `mgcv`, which tend to underestimate p-values when smoothing parameters are estimated. As we will show in Monte Carlo simulations in Section 2.6.3, the proposed test not only performs competitively compared to restricted likelihood ratio tests (RLRT, see e.g. Crainiceanu, Ruppert, Claeskens & Wand, 2005),but also allows to incorporate spatially adaptive smoothed curves without any additional effort.

2.6 Monte Carlo Studies

2.6.1 Simulation 1: Simultaneous Confidence Bands for Additive Models

First, we generate data from model (2.1) for $d = 3$ with homogeneous functions and i.i.d. Gaussian errors. The covariates are taken to be independent and uniformly distributed over $[0,1]$. The true functions f_j, shown in Figure 2.1(a) – (c) (centered to have zero mean), are simulated according to

$$\begin{aligned} f_1(x) &= \sin^2\{2\pi(x-0.5)\}, \\ f_2(x) &= \frac{6}{10}\beta_{30,17}(x) + \frac{4}{10}\beta_{3,11}(x), \\ f_{31}(x) &= x(1-x), \end{aligned}$$

with $\beta_{l,m} = \Gamma(l+m)\{\Gamma(l)\Gamma(m)\}^{-1}x^{l-1}(1-x)^{m-1}$. Functions f_1 and f_2 were also considered in Krivobokova, Kneib & Claeskens (2010), while f_{31} was used by Claeskens & Van Keilegom (2003). We scaled all three functions such that their standard deviations are all equal to one providing comparable signal-to-noise ratios (SNR).
We consider three different sample sizes (300, 600 and 1000), $k_j = 40, \ j = 1,2,3$

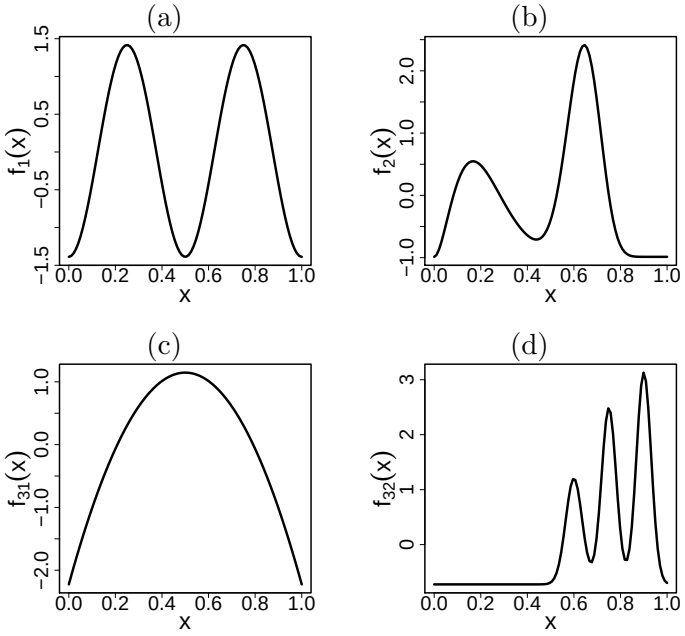

Figure 2.1: True functions in simulations 1 (top and bottom left) and 2 (top and bottom right) scaled to have variance 1.

knots and $\sigma \in \{0.33,\ 0.5,\ 1.0\}$, corresponding to medium, low and very low SNR, that is $\sqrt{\text{Var}\{f_j(x_j)\}}/\sigma \in \{3,\ 2,\ 1\}$. We used B-spline bases of degree three with penalties on the integrated squared second derivatives ($q = 2$) of the spline functions. Results for $k_j = 80$ knots were very similar and are therefore discarded. Table 2.1 shows the coverage rates based on a Monte Carlo sample size of 1000 and nominal coverage $100(1 - \alpha)\% = 95\%$. All coverage rates are very close to the nominal level of 0.95, except for f_2 in the case of $\sigma = 1.0$ and $n = 300$. In the latter case, the SNR is too low for the given small sample size such that the second peak of function f_2 could not be recovered frequently. This led to coverage rates lower than the nominal level, since the confidence bands were not correctly centered. Note, however, that this setting is very extreme compared to common settings used in simulations to test the performance of other approaches to simultaneous confidence bands (e.g. Claeskens & Van Keilegom, 2003), where usually considerably larger signal-to-noise ratios are used. Compared to these studies, we thus find that our approach works rather well also in quite unfavorable data situations.

Additionally, we replicated the simulation setting with covariates sampled randomly

Table 2.1: Coverage rates in simulations together with average areas in parenthesis. For simulation 2 results for either constant smoothing parameters and error variance (columns (i)) or varying error variance $\sigma^2(x_2)$ and adaptive smoothing parameter $\lambda_3(\tau_3)$ for f_{32} (columns (ii)) are given.

		Simulation 1			Simulation 2					
σ	n	f_1	f_2	f_{31}	f_1		f_2		f_{32}	
					(i)	(ii)	(i)	(ii)	(i)	(ii)
0.33	300	0.94	0.94	0.95	0.93	0.94	0.89	0.94	0.92	0.93
		(0.45)	(0.49)	(0.31)	(0.46)	(0.44)	(0.51)	(0.51)	(0.69)	(0.51)
	600	0.95	0.94	0.95	0.95	0.95	0.89	0.94	0.93	0.95
		(0.35)	(0.38)	(0.23)	(0.36)	(0.34)	(0.39)	(0.38)	(0.52)	(0.38)
	1000	0.96	0.95	0.96	0.95	0.94	0.88	0.95	0.94	0.95
		(0.28)	(0.31)	(0.19)	(0.29)	(0.27)	(0.32)	(0.31)	(0.42)	(0.29)
0.50	300	0.94	0.93	0.94	0.93	0.95	0.90	0.94	0.90	0.92
		(0.61)	(0.67)	(0.42)	(0.63)	(0.62)	(0.70)	(0.69)	(0.95)	(0.73)
	600	0.94	0.95	0.95	0.95	0.94	0.91	0.95	0.92	0.95
		(0.48)	(0.52)	(0.32)	(0.48)	(0.47)	(0.53)	(0.53)	(0.72)	(0.55)
	1000	0.95	0.95	0.96	0.95	0.95	0.91	0.94	0.93	0.95
		(0.39)	(0.42)	(0.26)	(0.39)	(0.39)	(0.43)	(0.43)	(0.59)	(0.42)
1.00	300	0.93	0.88	0.95	0.91	0.94	0.87	0.90	0.71	0.81
		(1.03)	(1.12)	(0.71)	(1.05)	(1.06)	(1.16)	(1.18)	(1.54)	(1.26)
	600	0.95	0.93	0.96	0.95	0.94	0.92	0.93	0.86	0.92
		(0.8)	(0.87)	(0.54)	(0.81)	(0.81)	(0.89)	(0.89)	(1.22)	(0.97)
	1000	0.94	0.94	0.97	0.95	0.94	0.93	0.94	0.89	0.92
		(0.66)	(0.72)	(0.44)	(0.67)	(0.66)	(0.73)	(0.73)	(1.01)	(0.76)

from the uniform distribution as well as with correlations between the covariates and obtained almost identical coverage rates. For the latter, we replaced the covariates by $(x_{1i}, x_{2i}, x_{3i})^t = \Phi(Z_i)$ with $Z_i \sim N(0, \Sigma_z)$ where $\Sigma_z = (1-\rho)I_{3\times 3} + \rho 1_3 1_3^t$ such that $\rho \in \{0.3, 0.5, 0.7\}$ relates to the correlation between the covariates which are marginally uniform on $[0, 1]$. The results are given in Table 2.2.

Table 2.2: Coverage rates and (average areas) for Simulation 1 with correlated covariates.

		$\rho = 0.3$			$\rho = 0.5$			$\rho = 0.7$		
σ	n	f_1	f_2	f_{31}	f_1	f_2	f_{31}	f_1	f_2	f_{31}
0.33	300	0.96	0.94	0.96	0.94	0.94	0.95	0.94	0.94	0.95
		(0.46)	(0.5)	(0.31)	(0.46)	(0.5)	(0.32)	(0.48)	(0.52)	(0.34)
	600	0.96	0.95	0.95	0.93	0.94	0.95	0.95	0.94	0.95
		(0.35)	(0.38)	(0.24)	(0.35)	(0.38)	(0.24)	(0.36)	(0.39)	(0.26)
	1000	0.94	0.95	0.96	0.95	0.96	0.96	0.96	0.96	0.96
		(0.28)	(0.31)	(0.19)	(0.29)	(0.31)	(0.2)	(0.3)	(0.32)	(0.21)
0.50	300	0.96	0.93	0.96	0.94	0.93	0.96	0.95	0.93	0.95
		(0.63)	(0.68)	(0.43)	(0.64)	(0.69)	(0.44)	(0.67)	(0.72)	(0.48)
	600	0.96	0.94	0.95	0.94	0.93	0.96	0.95	0.93	0.95
		(0.48)	(0.52)	(0.32)	(0.48)	(0.53)	(0.33)	(0.5)	(0.54)	(0.36)
	1000	0.94	0.95	0.96	0.95	0.95	0.96	0.96	0.96	0.96
		(0.39)	(0.43)	(0.26)	(0.4)	(0.43)	(0.27)	(0.41)	(0.45)	(0.29)
1.00	300	0.94	0.88	0.95	0.93	0.88	0.95	0.94	0.89	0.95
		(1.06)	(1.14)	(0.72)	(1.08)	(1.16)	(0.75)	(1.15)	(1.23)	(0.83)
	600	0.96	0.92	0.94	0.94	0.91	0.95	0.95	0.92	0.95
		(0.81)	(0.88)	(0.55)	(0.82)	(0.89)	(0.57)	(0.87)	(0.93)	(0.62)
	1000	0.94	0.94	0.96	0.95	0.94	0.96	0.95	0.95	0.95
		(0.66)	(0.72)	(0.45)	(0.68)	(0.73)	(0.46)	(0.71)	(0.76)	(0.5)

2.6.2 Simulation 2: Additive Model with Locally Adaptive Smoothed Components and Heteroscedasticity

In the second simulation study, function f_{31} of simulation 1 is replaced by function f_{32} shown in Figure 2.1(d) which is defined as

$$f_{32}(x) = \exp\{-400(x-0.6)^2\} + \frac{5}{3}\exp\{-500(x-0.75)^2\} + 2\exp\{-500(x-0.9)^2\}.$$

This function was also considered e.g. in Krivobokova, Crainiceanu & Kauermann (2008) and exhibits strong heterogeneity. Further, we introduce heteroscedasticity by specifying $\sigma(x_2) = \sigma - 0.2(x_2 - \overline{x}_2)$ where $\bar{x}_2$ denotes the arithmetic mean $\bar{x}_2 = n^{-1}\sum_{i=1}^{n} x_{2i}$. We consider either (i) constant smoothing parameters and error variance or (ii) varying error variance $\sigma^2(x_2)$ and adaptive smoothing parameter $\lambda_3(\tau_3)$ for f_{32} ($k_{w_3} = k_v = 5$ knots). All other settings remain the same as in Section 2.6.1.

Table 2.1 shows the coverage rates for $100(1-\alpha)\% = 95\%$. Coverage probabilities for function f_1 are very close to the nominal level regardless whether heterogeneities are taken into account or not except for $\sigma = 1,\ n = 300$ where the apparently worse overall model fit in (i) led to undercoverage. For function f_2 coverage probabilities improve considerably by taking heteroscedasticity into account such that rates of 0.94 or 0.95 are achieved except for the $\sigma = 1,\ n = 300$ case. Note the virtually identical average areas in (i) and (ii), i.e. the improvement is not ascribed to overall wider confidence bands. Locally adaptive estimation of f_{32} leads to a similar improvement and nearly perfect coverage rates were obtained, except for $n = 300$ and the very low SNR. Further, the average sizes of the bands are decreased notably, due to improved estimation of the horizontal part of f_{32}. However, estimation of the wiggly part of function f_{32} regularly failed for the smallest sample size or high noise settings, resulting in slight undercoverage in these cases. That is, although the volume-of-tube formula does not require $n \to \infty$, we observe improved coverage probabilities for increasing sample sizes, due to more precise function estimation.
Summarizing, the sample size must be large enough in low signal-to-noise settings such that the functions can properly be recovered, which is, however, a feature common to all approaches to confidence bands. Overall, we found the approach to perform very well even in these relatively complex models and extreme settings.

2.6.3 Simulation 3: Nonparametric Specification Test

We now compare the performance of the proposed test with the restricted likelihood ratio test of Crainiceanu, Ruppert, Claeskens & Wand (2005). We consider additive models with i. i. d. Gaussian errors

$$\begin{aligned} Y &= \mu_j(x_1, x_2, x_3) + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, \sigma^2 I),\ j = 1, 2, 3 \text{ with} \\ \mu_1(x_1, x_2, x_3) &= \varphi f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + f_{32}(x_3) \\ \mu_2(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + \varphi f_2(x_2) + x_3 + f_{32}(x_3) \\ \mu_3(x_1, x_2, x_3) &= f_1(x_1) + x_2(1 - x_2) + f_2(x_2) + x_3 + \varphi f_{32}(x_3) \end{aligned}$$

where $\varphi \in [0; 0.6]$ corresponds to the separation distance between the null and the alternative. We test for no effect, second degree polynomial and for linearity of the components $f_1^*(x_1) = \varphi f_1(x_1)$, $f_2^*(x_2) = x_2(1-x_2)+\varphi f_2(x_2)$ and $f_3^*(x_3) = x_3+\varphi f_{32}(x_3)$, respectively. To do so, B-spline bases with $(p = 1,\ q = 1)$, $(p = 5,\ q = 3)$ and $(p = 3,\ q = 2)$, respectively, are used.
Further, we choose $\sigma = 0.33$, $n = 300$, $k_j = 40,\ j = 1, 2, 3$ and $k_{w_3} = 5$. (Results for

$n = 600$ led to the same conclusions and are therefore not reported here.) Three Monte Carlo simulations with 1000 replications each were carried out.
Critical values for the RLRT test were computed using the simulation based approximation to the RLRT distribution implemented in the R package `RLRsim` (see Scheipl, Greven & Küchenhoff, 2008 which also includes a comprehensive comparisons of RLRT with F-type tests). The power curves of the proposed test and the RLRT test are virtually identical. The rejection rates are given in Figure 2.2.

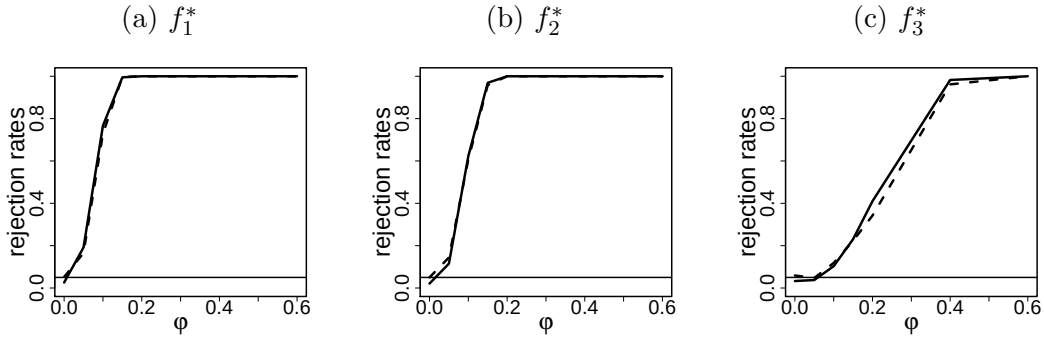

Figure 2.2: Empirical power curves of the proposed test (solid lines) and RLRT test (dashed lines) in simulation 3.

2.7 Studying Undernutrition in Kenya

We start by estimating the model (2.5). Figure 2.3 shows the estimated functions based on B-splines with $p = 5$, $q = 3$ and $k_1 = 40$, $k_2 = k_3 = 30$. In Figure 2.3(b), the partial residuals seem to exhibit a larger variability for small BMI values than for large BMI values. This could indicate a dependency between the Body Mass Index and the variance of the error term, which we want to explore by modeling the error variance as a smooth function of bmi. Further, the bump between the ages of 30 and 50 months in the enlarged plot of $\hat{f}_1(age)$ shown as grey line in Figure 2.4(a) could be an artefact due to the constant smoothing parameter used. Since the Z-score decreases rapidly in the first 20 months and remains nearly constant afterwards, it seems reasonable to estimate the effect of age with a locally adaptive smoothing parameter, as discussed in Section 2.4.3. Note that WHO (2006) also faced this problem of functional heterogeneity in their derivation of reference standards used to construct the Z-scores. However, instead of locally adaptive smoothing, a rather crude approach was chosen to address the issue (see Section 2.3). Naturally, neglecting these heterogeneities in smoothness and error

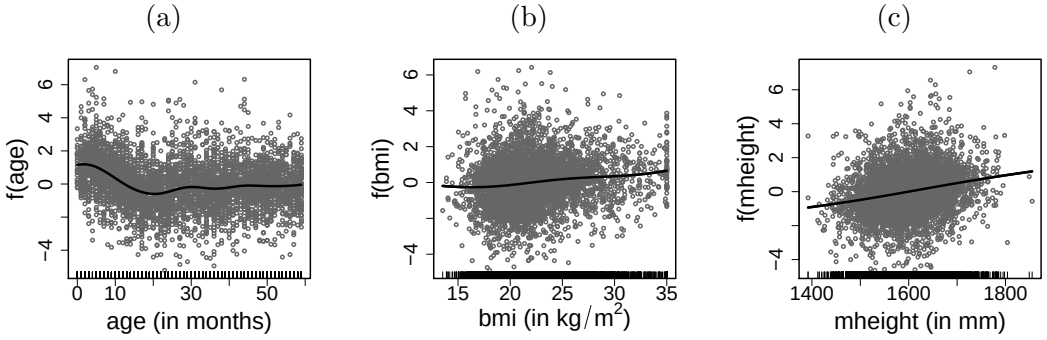

Figure 2.3: Estimated effects with corresponding partial residuals in model (2.5).

variance could lead to wrong conclusions. Figure 2.4 shows the results for model (2.5) supplemented by these two features. For completeness, parametric estimates are given in Table 2.3.

The density plot in Figure 2.5(a) shows that now the residual distribution is reasonably close to the Gaussian distribution. That is, we can consider the distributional assumption for the validity of the given confidence bands to be fulfilled. The estimated smoothing parameter function $\hat{\lambda}_1(\tau_1)$ shown in Figure 2.5(b) penalizes the roughness of $f_1(age)$ more strongly for larger age values. This ensures that the 'wiggliness' of the age effect between 30 and 50 month disappears. The estimated function of the residual standard deviation in Figure 2.5(c) indicates a slightly decreasing trend with *bmi*, however, barely affecting the width of the confidence bands in Figure 2.4.

The resulting estimated fit of the mother's BMI is positive and statistically significant based on a 5% significance level, since the zero line lies not entirely inside the simultaneous confidence band. However, the effect of *bmi* is more or less linear and, according to the test proposed in Section 2.5, does not significantly deviate from the parametric fit (with a p-value of 0.652). That is, the inverted U shape of the effect of the mother's BMI mentioned before is not confirmed for our Kenyan data. Similarly, the estimated function of the mother's height (*mheight*) is virtually linear and does not significantly deviate from the parametric fit (p-value 1). Regarding the age effect, we find a clearly nonlinear relationship and a significant deviation from the parametric linear fit (the p-value of our test is < 0.0001). Note also that the hypothesis of a quadratic age effect would be rejected indicating that the commonly used parametric models quadratic in *age* are vulnerable to misspecification bias and inference for other variables of interest could be misleading.

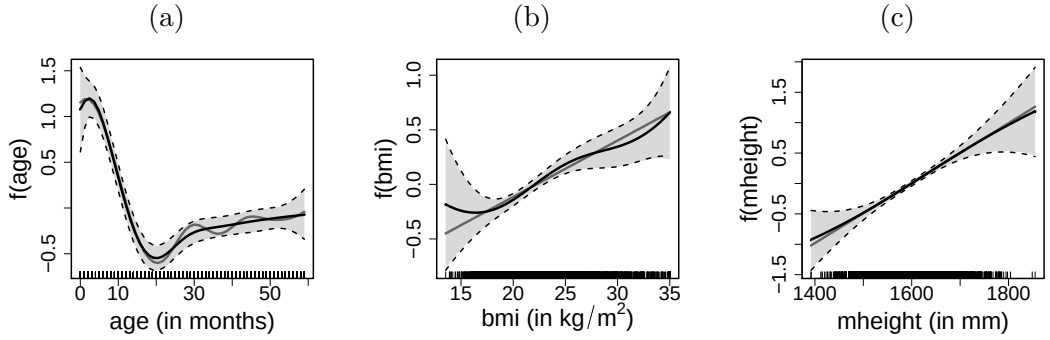

Figure 2.4: Estimated effects in the final model. As gray lines, in (a) the fit assuming constant smoothing parameter and in (b) and (c) the linear fits are superimposed. 95% simultaneous confidence bands assuming homoscedasticity (light gray area) and heteroscedasticity (dashed lines) are practically identical and can therefore hardly be distinguished.

The child's nutritional status seems to be more or less constant for the first three or four months of age, which is, however, associated with high uncertainty. Then, as already been suggested by the nutritional literature, there is a virtually linear deterioration until some inflection point at about 20 months of age after which there seems to be some improvement. In order to investigate whether this catching-up is real, i.e. statistically significant, we compute the first derivative of the function of age, which is given in Figure 2.6. The slope observed after the inception point until approximately 28 months is marginally significant on a 5% level, since the 95% confidence band around the first derivative does not include the zero line for this range. Afterwards, the zero line is included, meaning that we cannot reject the null hypothesis of no catching-up for ages larger than 28 months.

Despite the efforts of the WHO to improve the comparability of Z-scores by age by introducing a new reference standard derived from samples of comparable populations of children younger and older than 24 months, there still could be some problems. For example, the estimated derivative could also be picking up the fact that children younger than 2 years were measured recumbent and children older than 2 years were measured standing. To account for this, the reference standard was adjusted assuming a difference of 0.7 cm as described in Section 2.3. Note that this difference is associated with high uncertainty. Also, using the mean differences between recumbent length and height instead of the median could make the estimate sensitive to a very likely skewed distribution. If this estimate is not appropriate or this difference is smaller for the Kenyan

Table 2.3: Parametric estimates. Reference category for regional effects is Nairobi province.

	Estimate	Std. Error	t value	p-value
(Intercept)	-1.380	0.091	-15.090	0.000
yearsofedu	0.041	0.006	6.778	0.000
rural	-0.100	0.060	-1.658	0.097
female	0.197	0.041	4.868	0.000
central	-0.132	0.101	-1.300	0.194
coast	-0.071	0.102	-0.700	0.484
eastern	-0.163	0.106	-1.539	0.124
nyanza	-0.129	0.102	-1.275	0.202
rift valley	-0.151	0.098	-1.553	0.121
western	-0.248	0.099	-2.495	0.013
north eastern	0.547	0.127	4.316	0.000

children (who were smaller in average than the sample of healthy children from well-to-do families which form the reference standard), this could have led to the observed effect which therefore has to be treated with caution. To see this, we show in Figure 2.6(b)-(c) what would happen if the difference between children measured lying down and standing was assumed to be only 0.3 cm. Given that the children in Kenya are generally much worse nourished than children in the reference standard, this might well be the case. As shown in the figure, if the difference were only 0.3 cm, the significant effect of catch-up growth would disappear. Similarly, if there is substantial age misreporting around that age group, the reliability of the finding of catch-up growth could be open to question.

2.8 Discussion

In this chapter we construct simultaneous confidence bands for additive models with varying residual variance and spatially heterogenous smooth components. In doing so, the use of the mixed model representation of penalized splines not only allows for the fast and efficient estimation of such complex models, it also helps to build simultaneous confidence bands with very good small sample properties instantly, that is without using bootstrap or other numerically demanding techniques. Moreover, this technique can be used to construct specification tests for the additive components. Our simulation study confirmed that the resulted coverage probabilities are very close to the nominal level even for small sample sizes and the specification test is competitive to simulation based

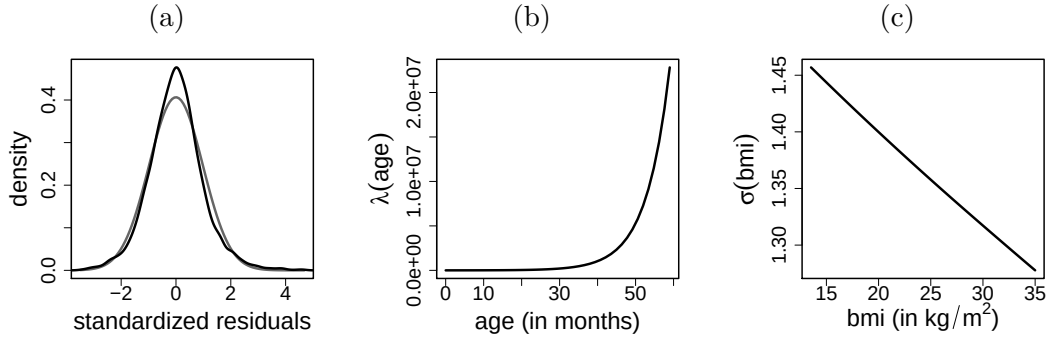

Figure 2.5: In (a), the gray line corresponds to the standard normal pdf. In (b) and (c), the estimated smoothing parameter function $\hat{\lambda}_1(\tau_1)$ and the estimated residual standard deviation $\hat{\sigma}(bmi)$ based on $k_v = k_{w_1} = 5$ knots are given.

alternatives. When studying data on undernutrition of children in Kenya the suggested model, the simultaneous confidence bands, and corresponding specification tests generated useful insights into drivers of undernutrition of Kenyan children, particularly the highly non-linear age affect. Our analysis indicates a statistically significant improvement of the stunting score between ages of 23 and 28 months. This, however, could also be due to differences in height measurements of children younger/older than 24 months and therefore requires further investigation. For children older than 28 months, no evidence for catch-up growth with respect to the reference population is found. From a model selection point of view, our analysis emphasizes the importance of flexible estimation of the age effect in order to avoid misspecification bias in the fully parametric models that are frequently employed in this context. Note that the data exhibit both heterogeneity in the functional form of some additive components as well as heteroscedasticity.
Possible further extensions are to include random effects and multidimensional components into the additive model, as well as to account for possible serial correlations in the data. It is important to note that the confidence bands considered rely explicitly on the assumption of normality of the data. Even though for symmetric distributions and sufficiently large sample sizes this assumption is less crucial and good results are typically obtained (see Loader & Sun, 1997), some corrections would be needed for highly skewed data. The proposed approach is quite fast and can readily be applied to large data sets despite its nonparametric nature. It is implemented in the R package *AdaptFitOS* described in Section 5.1.

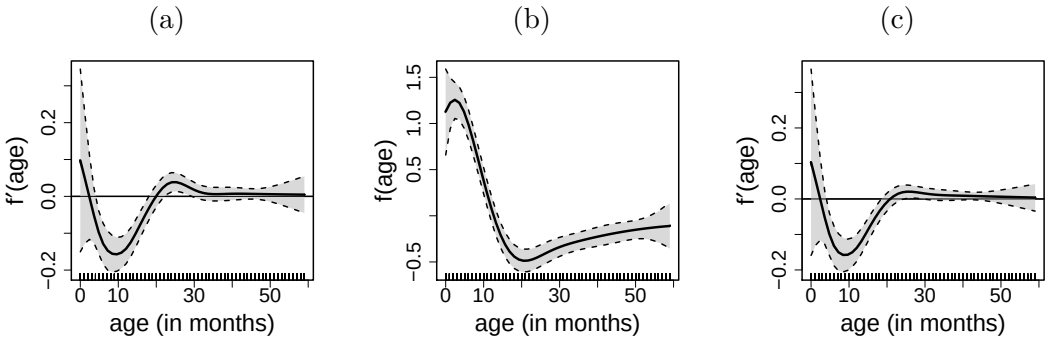

Figure 2.6: (a): Estimated first derivative of the age effect with 95% simultaneous confidence band. (b) and (c): Estimated age effect and its first derivative assuming that the recumbent length and standing height only differ by 0.3 cm.

3 Bayesian Geoadditive Sample Selection Models

Abstract: *Sample selection models attempt to correct for the presence of non-randomly selected data in a two-model hierarchy where, on the first level, a binary selection equation determines whether a particular observation will be available for the second level, i.e. in the outcome equation. Ignoring the non-random selection mechanism induced by the selection equation may result in biased estimation of the coefficients in the outcome equation. In the application that motivated this research, we analyze relief supply in earthquake affected communities in Pakistan, where the decision to deliver goods represents the dependent variable in the selection equation while factors that determine the amount of goods supplied are analyzed in the outcome equation. In this application, the inclusion of spatial effects is necessary since the available covariate information on the community level is rather scarce. Moreover, the high temporal dynamics underlying the immediate delivery of relief supply after a natural disaster calls for nonlinear, time-varying effects. We propose a geoadditive sample selection model that allows us to address these issues in a general Bayesian framework with inference being based on Markov chain Monte Carlo simulation techniques. The proposed model is studied in simulations and applied to the relief supply data from Pakistan.*

3.1 Introduction

A phenomenon frequently occurring in practice is non-randomly selected data with possibly severe impact on parameter estimates derived from statistical models ignoring this sample selection. In the application that motivated our research (see Benini, Conley, Dittemore & Waksman (2009) for a detailed introduction), we are faced with sample selection in a data set on relief supply. On 8 October 2005, an earthquake struck the northern part of Pakistan and Indian Kashmir, affecting a population of about 3.5 million

people. Though national and international delivery of relief supply started immediately, the distribution in the earthquake affected area was restricted, mainly due to constraints in transport capacities both for road and air transport. As a consequence, not all requests for relief supply could be satisfied but only a selected subset. We are interested in analysing both the factors that drive the decision to deliver relief supply after a specific request and the factors that determine the actual amount of delivered goods. Since it is very likely that correlations between the probability of positive decisions and delivered amounts will be present, it is important to avoid the introduction of sample selection bias by analysing both quantities simultaneously. Moreover, our application calls for flexible extensions of standard, parametric sample selection models (as applied to the same data in Benini, Conley, Dittemore & Waksman, 2009). Our database consists of delivery requests and actual deliveries for 87 Union Councils on 199 days. As a consequence, time-varying effects as well as spatial effects induced by unobserved spatially varying covariates should be included in a thorough analysis. We will therefore introduce geoadditive sample selection models and Bayesian inferential schemes based on Markov chain Monte Carlo (MCMC) simulation. Note that the structure of our data with a low number of observations corresponding to positive amounts delivered and a high number of zero deliveries, may also be modeled in different contexts. Zero-inflated models and two-part models are such alternatives (see Min & Agresti (2002) for a survey). However, unlike the sample selection model, their standard formulations do not include correlations between the two processes which is a crucial assumption in our reasoning. Therefore, we will formulate our model in the context of sample selection models in the following.

Reflecting the two-stage mechanism underlying the selected sampling process, the classical sample selection model consists of two model equations. The *selection equation* is formulated in terms of a binary probit model

$$P(y_{i1}^* = 1) = \Phi(\eta_{i1}), \quad i = 1, \ldots, n,$$

where the binary indicator y_{i1}^* indicates whether observation i is selected ($y_{i1}^* = 1$) or not ($y_{i1}^* = 0$), Φ is the standard normal cumulative distribution function and η_{i1} is a predictor formed of covariates. In our application, $y_{i1}^* = 1$ relates to a positive decision to deliver relief supply and η_{i1} is correspondingly combined from covariates influencing this decision.

The *outcome equation* defines a Gaussian linear model for those observations that have been selected in the first place, i.e.

$$y_{i2} = \eta_{i2} + \varepsilon_{i2} \qquad \text{observed only if } y_{i1}^* = 1, \tag{3.1}$$

where y_{i2} is a real-valued response variable, η_{i2} is a second predictor combination of covariates, and $\varepsilon_{i2} \sim \mathrm{N}(0, \sigma_2^2)$ are random errors. Often, the sample selection model is also defined in such a way that y_{i2} is equal to zero instead of unobserved if $y_{i1}^* = 0$. This interpretation in some sense fits better to our application (where y_{i2} will be the amount of goods delivered upon a request) than the classical definition (3.1) and also provides a connection to zero-inflated models.
It is often plausible to assume correlations between the response variables of the two equations. For example, in our analysis it will turn out that a positive decision to deliver is associated with smaller amounts delivered. Such correlations can be included into the model formulation when considering the latent Gaussian model representation of the probit model where a linear model

$$y_{i1} = \eta_{i1} + \varepsilon_{i1}, \quad \varepsilon_{i1} \sim \mathrm{N}(0, 1)$$

is assumed for the latent response y_{i1} and

$$y_{i1}^* = 1 \quad \Leftrightarrow \quad y_{i1} \geq 0.$$

The principal idea behind this formulation is to consider y_{i1} as a latent variable generally interpreted as some kind of utility associated with $y_{i1}^* = 1$. In our application, y_{i1} may be interpreted as a continuous score that is assigned to a specific request for relief supply and determines whether goods will be delivered. This score will be determined by different influential factors such as the urgency of the request but also availability of the required resources. The latent Gaussian representation now allows to correlate selection and outcome equation by assuming a correlated bivariate normal distribution for the error terms, i.e.

$$\begin{pmatrix} \varepsilon_{i1} \\ \varepsilon_{i2} \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_1^2 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix} \right). \tag{3.2}$$

In addition, the latent formulation of the probit model also facilitates Bayesian inference where the imputation of the unobserved latent variables y_{i1} yields simple Gibbs sampling steps and avoids the necessity to derive suitable proposal densities in a Metropolis Hastings sampler.
Since their introduction by Heckman (1979), sample selection models have been heavily employed in particular in the econometric literature but also in the social sciences (see for example Winship & Mare, 1992, or Sigelman & Zeng, 1999). Most of these papers considered parametric sample selection models where the predictors are formed as linear combinations of covariates, i.e. $\eta_{ij} = u_{ij}^t \gamma_j$, where u_{ij} and γ_j are a vector of covariates and a corresponding vector of regression coefficients for either selection ($j = 1$) or out-

come ($j = 2$) equation. Especially if some or all of the covariates in the selection and the outcome predictor are the same, severe consequences have to be expected when ignoring non-random selection in the outcome equation. Estimation in parametric sample selection models is typically based on the two-step estimation procedure proposed by Heckman (1979). Based on estimates for the selection equation, a correction component (the inverse Mills ratio) is added to the outcome equation to obtain valid estimates. The two-step estimates require that the model specifications for selection and outcome equations are different, i.e. at least one covariate has to be excluded from the outcome equation and this is referred to as an exclusion restriction in the literature. Our simulations indicate that estimates obtained by the Bayesian approach considered in this chapter can still be reliable (at least in case of symmetric error distributions) when no exclusion restriction is available and two-step estimation gets increasingly instable.
In our application, a parametric model is deemed insufficient for several reasons. First of all, the data have been collected over time and besides a general temporal change in both the frequency and amount of deliveries, it is also expected that covariate effects are changing over time. This reflects, for example, the varying impact of transport capacity limitations or changing knowledge about the requirements for relief supply. Such temporal changes in covariate effects can be addressed in the framework of varying coefficient models (Hastie & Tibshirani, 1993) requiring nonparametric modeling strategies for the temporal effects. Moreover, the covariate database may be expected to miss important covariates, at least some of which follow a spatial pattern. This results in spatially correlated data and can (at least partly) be accounted for by including a spatial effect. Consequently, we consider predictors of the form

$$\eta_{ij} = u_{ij}^t \gamma_j + x_{ij1} g_{j1}(t) + \ldots + x_{ijp} g_{jp}(t) + f_{j,\text{spat}}(s_i)$$

in our application, where $u_{ij}^t \gamma_j$ corresponds to usual parametric effects, $g_{j1}(t),\ldots,g_{jp}(t)$ are time-varying effects of covariates $x_{ij1},\ldots,x_{ijp}$, and $f_{j,\text{spat}}(s_i)$ is a spatial effect of a regional variable s_i. While most of the literature on semiparametric sample selection models focusses on relaxing the distributional assumption on the error terms (see Vella (1998) or Lee (2000) for overviews), we are interested in making the predictor equation more flexible. Das, Newey & Vella (2003) consider the estimation of flexible, nonlinear effects and extend the two-step estimation procedure to this situation. Chib, Greenberg & Jeliazkov (2009) propose a Bayesian estimation scheme also for sample selection models with flexible nonlinear effects. The latter are modeled through Bayesian versions of smoothing splines and estimation is based on Markov chain Monte Carlo simulation techniques. We will further extend this approach to a Bayesian estimation scheme based

on low-rank penalized splines for nonlinear effects, varying coefficient terms and Markov random field priors for spatial effects.
The rest of this chapter is organized as follows: Section 3.2 systematically introduces geoadditive sample selection models within a unifying framework. Section 3.3 describes Bayesian inference and the associated MCMC sampling steps. The derived methodology is validated in simulation studies in Section 3.4 and applied to the relief supply data in Section 3.5. The final Section 3.6 provides comments on possible extensions and directions of future research.

3.2 Geoadditive Sample Selection Models

The most general sample selection model that will be relevant for our work is defined by predictors

$$\eta_{ij} = u_{ij}^t\gamma_j + f_{j1}(z_{ij1}) + \ldots + f_{jq}(z_{ijq}) + x_{ij1}g_{j1}(z_{ij,q+1}) + \ldots + x_{ijp}g_{jp}(z_{ij,q+p}) + f_{j,\text{spat}}(s_i), \quad j = 1, 2,$$

that extend the model considered in the introduction by including nonparametric effects $f_{j1}(z_{ij1}),\ldots,f_{jq}(z_{ijq})$ of continuous covariates $z_{ij1}, \ldots, z_{ijq}$ and also admit continuous effect modifiers $z_{ij,q+1}, \ldots, z_{ij,q+p}$ other than time t. Of course, in practice the predictor specifications for selection and outcome equation do not have to be the same and in particular will in general not contain the same number of nonparametric effects or varying coefficient terms. However, to ease notation, we will suppress this in the following.

3.2.1 Parametric Effects

For parametric effects γ_j, we assume flat, noninformative priors $p(\gamma_j) \propto \text{const}$ throughout this chapter. This assumption could easily be replaced by informative Gaussian prior distributions but in the absence of further prior knowledge, we prefer the noninformative prior choice that avoids specification of hyperparameters.

3.2.2 Nonparametric Effects

To obtain a low-rank representation with relatively few parameters for the nonparametric effects, we adopt the Bayesian P-spline specification introduced by Lang & Brezger (2004). The idea builds on the frequentist penalized spline approach popularized by Eilers & Marx (1996), where each of the nonparametric effects $f(z)$ (dropping indices for the sake of simplicity) is approximated by a B-spline basis $B_1(z), \ldots, B_K(z)$, of degree

D, i.e.

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z).$$

While the degree D of the spline basis can typically be chosen according to subject matter considerations about the differentiability of $f(z)$, the number of basis functions K is harder to determine. A large number of basis functions yields a very flexible basis, but is prone to overfitting the data. On the other hand, choosing a low-dimensional basis risks missing important features in the functional form of $f(z)$. As a remedy, penalized splines are built upon a moderately sized basis with 20 to 40 basis functions as a suitable default choice, but add a penalty term to the estimation criterion. In the approach of Eilers & Marx (1996), simple squared differences of the basis coefficients are shown to approximate the integrated squared derivative penalty well-known from smoothing splines.

From a Bayesian perspective, adding a penalty to the likelihood corresponds to assigning an informative prior distribution to the basis function coefficients $\beta = (\beta_1, \ldots, \beta_K)^t$. To be more specific, the difference penalty corresponds to a random walk (RW) assumption, with

$$\beta_k = \beta_{k-1} + u_k, \qquad \text{and} \qquad \beta_k = 2\beta_{k-1} - \beta_{k-2} + u_k$$

for first and second order random walks, Gaussian innovations u_k i.i.d. $N(0, \tau^2)$, and noninformative priors for the initial parameters. The variance of the random walk acts as a smoothing parameter that governs the trade off between fidelity to the data (τ^2 large) and smoothness of the function estimate (τ^2 small).

The joint prior distribution for the coefficient vector β can be shown to be a multivariate Gaussian distribution of the form

$$p(\beta|\tau^2) \propto \left(\frac{1}{2\tau^2}\right)^{\frac{\text{rank}(\Delta)}{2}} \exp\left(-\frac{1}{2\tau^2}\beta^t \Delta \beta\right). \tag{3.3}$$

The penalty or precision matrix Δ is given by the cross-product of a difference matrix of appropriate order, i.e. $\Delta = \mathcal{D}^t\mathcal{D}$. Due to the noninformative prior for the initial parameters, a polynomial of order $d-1$ remains unpenalized by a d-th order random walk. As a consequence, the joint prior distribution is partially improper, reflected in the fact that Δ is rank-deficient.

The vector of function evaluations $f = (f(z_1), \ldots, f(z_n))^t$ can be written as $f = Z\beta$, where Z contains the evaluations of the basis functions.

3.2.3 Varying Coefficient Terms

Penalized splines are also useful in the context of varying coefficient terms $xg(z)$, where the effect of x is varying smoothly over the domain of z (Hastie & Tibshirani, 1993). Since $g(z)$ is assumed to be a smooth function of z, we can again apply penalized splines for their estimation. As a consequence, the vector of function evaluations g is again given by $Z\beta$. When considering the vector of contributions to the predictor, i.e. $g^* = (x_1 g(z_1), \ldots, x_n g(z_n))^t$, the matrix Z has to be multiplied row-wise with the values of the interaction variable leading to

$$g^* = \text{diag}(x_1, \ldots, x_n) Z\beta = Z^*\beta$$

where $Z^* = \text{diag}(x_1, \ldots, x_n)Z$. Again, a random walk prior can be assigned to the vectors of regression coefficients.

3.2.4 Spatial Effects

In our application, we require a suitable prior distribution for spatial effects based on areal data. As a consequence, we require a prior that takes spatial closeness between areas into account. This can be conceptualized by considering a neighborhood structure for the areas and by defining a Markov random field prior based on this neighborhood structure (Rue & Held, 2005). We define two areas to be neighbors if they share a common boundary and assign separate coefficients β_s representing the spatial effect in region s.

The assumption of a Markov random field for the coefficient vector $\beta = (\beta_1, \ldots, \beta_S)^t$, where S denotes the number of areas, corresponds to the assumption that the effect of an area s is conditionally Gaussian, with the mean of the effects of neighboring areas as expectation and a variance that is inverse proportional to the number of its neighbors N_s:

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

where δ_s contains all neighbors of region s. From the conditional prior specification, the joint prior distribution can be derived and is again of the multivariate Gaussian form (3.3). The precision matrix is now given by an adjacency matrix that reflects the neighborhood structure underlying the areas. The vector of evaluations of the spatial function $f_{\text{spat}} = (f_{\text{spat}}(s_1), \ldots, f_{\text{spat}}(s_n))^t$ can again be written as $Z\beta$, where Z is an incidence matrix of zeros and ones that links each observation to the corresponding

spatial effect.

3.2.5 Generic Model Representation

In summary, we find the same structure for all effects contained in our geoadditive sample selection model: The vector of function evaluations can be written as the product of a design matrix and a possibly high-dimensional vector of regression coefficients. Combining all observations in the predictor vectors $\eta_j = (\eta_{1j}, \ldots, \eta_{n_j,j})^t$ with dimension n_j corresponding to the number of observations for selection and outcome equation, therefore allows us to introduce a general matrix-vector representation of the model. After appropriate re-indexing, we obtain the model equations

$$\eta_j = U_j\gamma_j + Z_{j1}\beta_{j1} + \ldots + Z_{jr}\beta_{jr}, \quad j = 1, 2,$$

where r denotes the overall number of nonparametric effects (smooth, varying coefficient or spatial) and U_j is a fixed effects design matrix. Similarly, all priors for nonparametric effects are multivariate Gaussian and can therefore be written as

$$p(\beta_{jl}|\tau_{jl}^2) \propto \left(\frac{1}{2\tau_{jl}^2}\right)^{\frac{\text{rank}(\Delta_{jl})}{2}} \exp\left(-\frac{1}{2\tau_{jl}^2}\beta_{jl}^t\Delta_{jl}\beta_{jl}\right), \quad l = 1, \ldots, r.$$

This very general structure will considerable facilitate the description of inferential procedures in the following section and is also extremely helpful when developing MCMC samplers that can be used regardless of the specific type of an effect.
The prior specification for nonparametric effects is completed by assigning a suitable hyperprior to the smoothing variance τ_{jl}^2. For the sake of convenience, we will consider conjugate inverse gamma priors $\tau_{jl}^2 \sim \text{IG}(a, b)$ throughout this chapter.

3.2.6 Priors for the Error Term Covariance Matrix

Finally, a suitable prior distribution has to be assigned to the covariance matrix of the error terms in (3.2). Since the variance of the selection equation is restricted to one, the standard choice of a conjugate inverse Wishart prior is not available. Instead, following Omori (2007) we consider a reparameterisation that allows to assign standard prior distributions of the free parameters. Therefore we write

$$\text{Cov}(\varepsilon_i) = \begin{pmatrix} \sigma_1^2 = 1 & \sigma_{12} \\ \sigma_{12} & \sigma_{2|1}^2 + \sigma_{12}^2 \end{pmatrix}$$

where $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2})^t$, $\sigma_{12} = \text{Cov}(\varepsilon_{i1}, \varepsilon_{i2})$, and $\sigma^2_{2|1} = \text{Var}(\varepsilon_{i2}|\varepsilon_{i1})$. In this parameterisation, a Gaussian prior can be assigned to the covariance, i.e. $\sigma_{12} \sim \text{N}(m_{\sigma_{12}}, s^2_{\sigma_{12}})$, while an inverse Gamma prior can be employed for the conditional variance $\sigma^2_{2|1}$, i.e. $\sigma^2_{2|1} \sim IG(a_{\sigma_{2|1}}, b_{\sigma_{2|1}})$. See Chib, Greenberg & Jeliazkov (2009) for a derivation of this prior specification based on an inverse Wishart prior for the covariance matrix.

3.3 Bayesian Inference

Employing the latent Gaussian formulation of the probit model (Albert & Chib, 1993) yields two Gaussian regression models with correlated error terms. After imputing the unobserved latent variables y_{i1}, the model definition therefore would equal a seemingly unrelated regression model, and Bayesian inferential schemes as developed in Lang, Adebayo, Fahrmeir & Steiner (2003) could in principle be used. However, due to sample selection, observations on the outcome equation are only available for parts of the observations. If all covariates of the outcome equation are also observed for the missing response variables, it is possible to impute also the missing response variables y_{i2} and to construct a complete data set in each MCMC iteration. This, in turn, then enables the application of methodology for seemingly unrelated regression, see for example Kai (1998) or van Hasselt (2005). However, we found in preliminary analyzes that this imputation approach typically shows mixing and convergence problems and, in particular, does not yield satisfactory estimates for the error covariance and therefore frequently fails to correct for the bias induced by sample selection.
We therefore follow Chib, Greenberg & Jeliazkov (2009) and Omori (2007) and consider a sampler that imputes latent Gaussian variables for the selection equation but uses only the observed responses from the outcome equation. Besides providing better estimation results, this also speeds up computation times since the imputation of unobserved outcomes is avoided. The full conditionals for all model parameters are then given as follows:

- The full conditionals for latent response y_{i1} are truncated normal

$$y_{i1}|\cdot \sim \begin{cases} \text{TN}_{(-\infty,0)}(\eta_{i1}, 1) & \text{if } y^*_{i1} = 0 \\ \text{TN}_{[0,\infty)}(m_{y_{i1}}, s^2_{y_{i1}}) & \text{if } y^*_{i1} = 1 \end{cases}$$

  where $\text{TN}_{[a,b]}(m, s^2)$ denotes a normal distribution with mean m and variance s^2

truncated to the interval $[a, b]$ and

$$\begin{aligned} m_{y_{i1}} &= \mathrm{E}(y_{i1}|y_{i2}, y_{i1}^* = 1) = \eta_{i1} + \frac{\sigma_{12}}{\sigma_{2|1}^2 + \sigma_{12}^2}(y_{i2} - \eta_{i2}), \\ s_{y_{i1}}^2 &= \mathrm{Var}(y_{i1}|y_{i2}, y_{i1}^* = 1) = 1 - \frac{\sigma_{12}^2}{\sigma_{2|1}^2 + \sigma_{12}^2} = \sigma_{1|2}^2. \end{aligned}$$

- The full conditionals for parametric effects γ_j are Gaussian $\gamma_j|\cdot \sim \mathrm{N}(m_{\gamma_j}, P_{\gamma_j}^{-1})$ with precision matrix

$$P_{\gamma_j} = \begin{cases} \left[\frac{1}{\sigma_{1|2}^2} U_1^t U_1\right]_{y_{i1}^*=1} + \left[U_1^t U_1\right]_{y_{i1}^*=0} & \text{if } j = 1 \\ \left[\frac{1}{\sigma_{2|1}^2} U_2^t U_2\right]_{y_{i1}^*=1} & \text{if } j = 2 \end{cases}$$

and mean

$$m_{\gamma_j} = \begin{cases} P_{\gamma_1}^{-1} \left(\left[\frac{1}{\sigma_{1|2}^2} U_1^t (y_1 - o_1)\right]_{y_{i1}^*=1} + \left[U_1^t (y_1 - \tilde{\eta}_1)\right]_{y_{i1}^*=0} \right) & \text{if } j = 1 \\ P_{\gamma_2}^{-1} \left(\left[\frac{1}{\sigma_{2|1}^2} U_2^t (y_2 - o_2)\right]_{y_{i1}^*=1} \right) & \text{if } j = 2 \end{cases}$$

where $[\ldots]_{y_{i1}^*=1}$ is used to denote that the matrices and vectors contained in the brackets are restricted to observations with $y_{i1}^* = 1$ (and analogously for $y_{i1}^* = 0$) and $\sigma_{1|2}^2$ denotes the conditional variance $\mathrm{Var}(\varepsilon_{i1}|\varepsilon_{i2})$ that was already involved in the full conditional for latent responses from the selection equation. The offset vectors o_j are given by

$$o_j = \begin{cases} \frac{\sigma_{12}}{\sigma_{2|1}^2 + \sigma_{12}^2} (y_2 - \eta_2) + [\tilde{\eta}_1]_{y_{i1}^*=1} & \text{if } j = 1 \\ [\sigma_{12}(y_1 - \eta_1)]_{y_{i1}^*=1} + [\tilde{\eta}_2]_{y_{i1}^*=1} & \text{if } j = 2 \end{cases}$$

and $\tilde{\eta}_j = \eta_j - U_j \gamma_j$ denotes the predictor vector excluding the parametric effects.

- The full conditionals for regression coefficients of nonparametric effects, varying coefficients and spatial effects are Gaussian $\beta_{jl}|\cdot \sim \mathrm{N}(m_{\beta_{jl}}, P_{\beta_{jl}}^{-1})$ with precision

matrix

$$P_{\beta_{jl}} = \begin{cases} \left[\frac{1}{\sigma^2_{1|2}} Z^t_{1l} Z_{1l}\right]_{y^*_{i1}=1} + \left[Z^t_{1l} Z_{1l}\right]_{y^*_{i1}=0} + \frac{1}{\tau^2_{1l}} \Delta_{1l} & \text{if } j = 1 \\ \left[\frac{1}{\sigma^2_{2|1}} Z^t_{2l} Z_{2l}\right]_{y^*_{i1}=1} + \frac{1}{\tau^2_{2l}} \Delta_{2l} & \text{if } j = 2 \end{cases}$$

and mean

$$m_{\beta_{jl}} = \begin{cases} P^{-1}_{\beta_{1l}} \left(\left[\frac{1}{\sigma^2_{1|2}} Z^t_{1l}(y_1 - o_1)\right]_{y^*_{i1}=1} + \left[Z^t_{1l}(y_1 - \tilde{\eta}_{1l})\right]_{y^*_{i1}=0} \right) & \text{if } j = 1 \\ P^{-1}_{\beta_{2l}} \left(\left[\frac{1}{\sigma^2_{2|1}} Z^t_{2l}(y_2 - o_2)\right]_{y^*_{i1}=1} \right) & \text{if } j = 2 \end{cases}$$

where the offset vectors o_j are given as for parametric effects and $\tilde{\eta}_{1l} = \eta_1 - Z_{1l}\beta_{l1}$ denotes the predictor of the selection equation excluding the l-th effect.

- The full conditionals for the smoothing variances are inverse gamma distributions $\tau^2_{jl}|\cdot \sim \text{IG}(\tilde{a}_{jl}, \tilde{b}_{jl})$ with parameters

$$\tilde{a}_{jl} = a + \frac{\text{rank}(\Delta_{jl})}{2}, \quad \tilde{b}_{jl} = b + \frac{1}{2}\beta^t_{jl}\Delta_{jl}\beta_{jl}.$$

- The full conditional for the error covariance is Gaussian $\sigma_{12}|\cdot \sim \text{N}(\tilde{m}_{\sigma_{12}}, \tilde{s}^2_{\sigma_{12}})$ with

$$\begin{aligned} \tilde{m}_{\sigma_{12}} &= \tilde{s}^2_{\sigma_{12}} \left(\frac{m_{\sigma_{12}}}{s^2_{\sigma_{12}}} + \left[\frac{1}{\sigma_{2|1}}(y_1 - \eta_1)^t(y_2 - \eta_2)\right]_{y^*_{i1}=1} \right) \\ \tilde{s}^2_{\sigma_{12}} &= \left(\frac{1}{s^2_{\sigma_{12}}} + \left[\frac{1}{\sigma_{2|1}}(y_1 - \eta_1)^t(y_1 - \eta_1)\right]_{y^*_{i1}=1} \right)^{-1}. \end{aligned}$$

- The full conditional for the conditional variance of the outcome equation is inverse gamma $\sigma^2_{2|1}|\cdot \sim \text{IG}(\tilde{a}_{\sigma_{2|1}}, \tilde{b}_{\sigma_{2|1}})$ with

$$\begin{aligned} \tilde{a}_{\sigma_{2|1}} &= a_{\sigma_{2|1}} + \frac{n_2}{2} \\ \tilde{b}_{\sigma_{2|1}} &= b_{\sigma_{2|1}} + \left([\sigma_{12}(y_1 - \eta_1)]_{y^*_{i1}=1} - (y_2 - \eta_2) \right)^t \left([\sigma_{12}(y_1 - \eta_1)]_{y^*_{i1}=1} - (y_2 - \eta_2) \right), \end{aligned}$$

where n_2 denotes the number of observations in the outcome equation.

Since all full conditionals reduce to well-known distributions, a Gibbs sampling scheme can be set up to perform Bayesian inference. We will use the mean of the posterior samples as an estimate for the posterior mean and will consider Bayesian credible intervals constructed from sample quantiles.
The two computational bottle necks of the sample selection Gibbs sampler are the generation of the latent Gaussian responses for the selection equation and the draws from high-dimensional Gaussian distributions to sample the parameter vectors β_{jl}. For drawing from truncated normals, we employed improved sampling schemes that do not rely on simple rejection sampling (Robert, 1995) but the corresponding simulation step still remains computationally demanding if the number of observations in the selection equation is high (as in our application). For drawing the regression coefficients β_{jl} we make use of sparse matrix algorithms that rely on the special structure of the precision matrix of the full conditionals (Lang & Brezger, 2004).

3.4 Simulations

3.4.1 Simulation Study 1: Parametric Sample Selection Models

In order to compare the proposed method with separate univariate regressions and Heckman models based on two-step estimation (computed by using package `sampleSelection` (Henningsen & Toomet, 2008) in `R`), a simulation with linear effects is conducted. Univariate regression estimates are calculated by maximum likelihood probit estimation and ordinary least squares estimation, respectively.
The model is specified through the predictors

$$\eta_{i1} = 2u_{i11} + u_{i12}, \qquad \eta_{i2} = 1.5u_{i21} + 2u_{i22}$$

All covariate values $(u_{i11}, u_{i21})^t$ and $(u_{i12}, u_{i22})^t$ are samples from bivariate Gaussian distributions with means 0.5, variances 1 and correlation ρ_{dm}. We examine correlated design matrices ($\rho_{dm} = 0.5$) and identical design matrices ($\rho_{dm} = 1.0$, i.e. $u_{i11} = u_{i12},\ u_{i12} = u_{i22}$). Further, bivariate Gaussian errors are considered with zero means, variances one and correlations $\rho_\varepsilon = 0.5$ and $\rho_\varepsilon = 0.9$, respectively. The simulation consists of 250 replications with 1000 observations each. According to the high amount of censoring in our application, approximately 95 percent of the total number of observations are censored in the first stage of the model such that only 50 observations remain in the outcome equation. In the case of the Bayesian sample selection model, the initial $5,000$ iterations are discarded (burn-in period) and from the subsequent $40,000$ iterations, every 40th iteration is recorded for inference. The high degree of thinning is applied to avoid pos-

			$\rho_{dm} = 0.5$			$\rho_{dm} = 1$		
ρ_ε		True	Univ.	SSM	2-step	Univ.	SSM	2-step
$\rho_\varepsilon = 0.5$	(Int Selection)	-5.5	-0.1803	-0.1985	-0.1803	-0.1007	-0.1189	-0.1007
	u_{11}	2.0	0.0653	0.0690	0.0653	0.0412	0.0418	0.0412
	u_{12}	1.0	0.0382	0.0427	0.0382	0.0169	0.0200	0.0169
	(Int Outcome)	0.0	0.4652	-0.0011	0.0158	1.9177	0.7989	0.2893
	u_{21}	1.5	-0.0422	0.0019	0.0018	-0.5321	-0.2198	-0.0791
	u_{22}	2.0	0.0047	0.0025	0.0035	-0.2561	-0.1157	-0.0501
$\rho_\varepsilon = 0.9$	(Int Selection)	-5.5	-0.1481	-0.1368	-0.1481	-0.1124	-0.1196	-0.1124
	u_{11}	2.0	0.0508	0.0453	0.0508	0.0467	0.0425	0.0467
	u_{12}	1.0	0.0318	0.0263	0.0318	0.0170	0.0157	0.0170
	(Int Outcome)	0.0	0.8343	-0.0312	0.0097	3.4446	0.4582	0.2649
	u_{21}	1.5	-0.0720	0.0046	0.0057	-0.9587	-0.1323	-0.0774
	u_{22}	2.0	0.0054	0.0044	0.0057	-0.4478	-0.0616	-0.0421

Table 3.1: Simulation study 1: Averaged estimation bias in the cases of correlated and identical design matrices. In the third column the true values are shown, while the other values are the difference of the averaged estimated values minus the true value.

sible sample autocorrelations. Nevertheless, sample autocorrelations of estimates in the selection equation (and to a lesser extent of the estimated components of the covariance matrix) do not completely disappear depending on the values of ρ_{dm} and ρ_ε. This is a well-known general issue in Bayesian (parametric) sample selection models (see Omori, 2007). However, since we did not observe consequences for the point estimates, we did not increase the given number of iterations.

Table 3.1 gives the estimation bias obtained by separate univariate regressions (Univ.), Bayesian sample selection model (SSM) and two-step estimation (2-step) averaged over the simulation runs. Table 3.2 shows empirical root mean squared errors (RMSE). Results for the estimated correlation between the errors and the variance of the error in the outcome equation σ_2^2 are given in Table 3.3.

The following conclusions can be drawn from the results of the simulation study (focusing on the outcome equation since in the selection equation estimation bias and mean squared errors are comparable in all methods):

- Using univariate regression, the estimation bias and RMSE increase with increasing correlations of the error terms and increasing correlations of the design matrices (Tables 3.1 and 3.2). In the case of $\rho_{dm} = 0.5$, considerable selection bias occurs only in the intercept, while in the case of identical design matrices all coefficients

are highly biased when using univariate regression. All sample selection models considerably reduce the estimation bias and mean squared errors in all settings.

- With increasing ρ_{dm}, both sample selection models increasingly underestimate ρ_ε and the associated selection bias (Table 3.3), i.e. the estimated coefficients get closer to those in univariate regression but are still less biased (Table 3.1).

- While the averaged estimation bias is lower for two-step estimation than for the Bayesian approach in the case of $\rho_{dm} = 1$, it is the other way round for the mean squared errors (Tables 3.1 and 3.2). Hence, two-step estimation appears to be less efficient (but less biased on average) than the Bayesian sample selection model in this case. In the case of low correlations of the design matrices, the differences are minimal. The two-step estimator is known to suffer from identification problems in the case of highly correlated design matrices resulting in instable estimates. The lower variability of the estimates in the Bayesian approach (reflected by the lower mean squared errors) might indicate that our approach is less prone to this issue.

- The RMSE of the estimated correlation is relatively high for both methods and in particular for identical design matrices. However, it is always lower in the Bayesian approach than in two-step estimation (Table 3.3).

- While the estimation bias for σ_2^2 only varies minimally over the different settings in the Bayesian approach, the bias is higher for the two-step estimator in the case of identical design matrices (Table 3.3). Regarding the RMSE of σ_2^2, there is an increase for both methods in the case of identical design matrices but to a much lesser extent in the Bayesian approach than in two-step estimation.

- In general, the value of ρ_{dm} has a higher impact on mean squared errors and estimation bias in all methods than the value of ρ_ε.

Additionally, we conducted a simulation study with lower degree of censoring (approx. 25% to 29%) which yielded similar results. In summary, the proposed Bayesian approach appears to be at least competitive to two-step estimation in the parametric setting and performs better in case of high correlations between the design matrices.

		$\rho_{dm} = 0.5$			$\rho_{dm} = 1$		
	Estimate	Univ.	SSM	2-step	Univ.	SSM	2-step
$\rho_\varepsilon = 0.5$	(Int Sel.)	0.6372	0.6568	0.6372	0.5874	0.5948	0.5874
	u_{11}	0.2716	0.2758	0.2716	0.2444	0.2465	0.2444
	u_{12}	0.1693	0.1698	0.1693	0.1626	0.1646	0.1626
	(Int Out.)	0.5415	0.3633	0.3479	2.0228	2.1439	2.7959
	u_{21}	0.1581	0.1489	0.1499	0.5826	0.6215	0.7974
	u_{22}	0.1163	0.1109	0.1117	0.3017	0.3068	0.3840
$\rho_\varepsilon = 0.9$	(Int Sel.)	0.6475	0.6240	0.6475	0.6020	0.5694	0.6020
	u_{11}	0.2719	0.2583	0.2719	0.2507	0.2355	0.2507
	u_{12}	0.1668	0.1544	0.1668	0.1590	0.1527	0.1590
	(Int Out.)	0.8715	0.2918	0.3049	3.4831	1.6861	2.5795
	u_{21}	0.1524	0.1152	0.1204	0.9784	0.4995	0.7359
	u_{22}	0.1068	0.0938	0.0961	0.4688	0.2467	0.3545

Table 3.2: Simulation study 1: Empirical root mean squared errors.

		$\rho_{dm} = 0.5$				$\rho_{dm} = 1$			
	Estimate	SSM		2-step		SSM		2-step	
		Bias	RMSE	Bias	RMSE	Bias	RMSE	Bias	RMSE
$\rho_\varepsilon = 0.5$	$\hat{\sigma}_2^2$	0.0365	0.2492	-0.0503	0.2307	-0.0305	0.3360	0.1658	0.5684
	$\hat{\rho}_\varepsilon$	-0.0365	0.1867	-0.0101	0.1921	-0.2568	0.4519	-0.1511	0.5726
$\rho_\varepsilon = 0.9$	$\hat{\sigma}_2^2$	0.0470	0.2477	-0.0375	0.2270	-0.0298	0.3958	0.1315	0.7569
	$\hat{\rho}_\varepsilon$	-0.0248	0.0924	-0.0006	0.1238	-0.1846	0.3541	-0.1634	0.4635

Table 3.3: Simulation study 1: Estimation bias (true $\sigma_2^2 = 1$) and root mean squared errors for the correlation between the errors and the variance in the outcome equation.

3.4.2 Simulation Study 2: Geoadditive Sample Selection Models

In this simulation study, the performance of the Bayesian semiparametric sample selection model is compared to separate univariate regressions based on generalized additive models. More precisely, results of the sample selection model are compared to an additive probit model in the selection equation on the one hand and to a Gaussian additive model in the outcome equation on the other hand. The estimation of the separate models is also Bayesian and carried out with the same sampling scheme as for the sample selection model but with the correlation of the errors fixed at zero. This allows all hyperparameters to be set equally, ensuring that the prior information is the same in both

methods.
The investigated model is specified through the predictors

$$\begin{aligned}\eta_{i1} &= f_{11}(x_{i1}) + f_{1,\text{spat}}(s_i)\\ \eta_{i2} &= f_{21}(x_{i1}) + f_{22}(x_{i2}) + f_{23}(x_{i3}) + f_{2,\text{spat}}(s_i).\end{aligned}$$

The included functions are given as follows:

$$\begin{aligned} f_{11}(x) = 2\Phi(x) - 1, \qquad & f_{21}(x) = 1 - \frac{1}{8}(x+2)^2,\\ f_{22}(x) = \sin(x) + 1.5 \cdot \exp(-10x^2), \qquad & f_{23}(x) = 1.5 \cdot \sin(\pi x)^2 \end{aligned}$$

where $\Phi(x)$ denotes the standard Gaussian distribution function. The functions $f_{j,\text{spat}}(s_i)$ are bivariate functions of the centroids of regions in a map of Baden-Württemberg and Bavaria shown in the top graphs of Figure 3.1, where the black colored regions in the left panel indicate a negative effect on selection, i.e. they are more likely to be censored. The covariate values are i.i.d. uniformly distributed

$$x_{i1} \sim U(-2, 2); \quad x_{i2} \sim U(-2, 2); \quad x_{i3} \sim U(0, 1).$$

Note that functions $f_{11}(x)$ and $f_{21}(x)$ as well as the spatial functions enter the model with the same covariates in selection and outcome equation. The error terms are i.i.d. bivariate Gaussian with zero means, variances $\text{Var}(\varepsilon_{i1}) = 1$ and $\text{Var}(\varepsilon_{i2}) = 2$ and correlation $\rho_\varepsilon = 0.9$. Again, 250 replications of the model each with $n = 500$ observations in the selection equation are simulated. Approximately 50% of the observations are censored. In all models, the first $5,000$ iterations are discarded and the $40,000$ following iterations are thinned by 40. The estimated nonparametric functions are based on cubic P-splines with 30 knots, second order random walk penalties and the choice $a = b = 0.001$ for the hyperparameters of variances.
Figure 3.2 shows posterior means of the smooth functions averaged over the simulation runs. Fits obtained by the Bayesian sample selection model (dashed lines) are compared to those obtained by univariate regression (dotted lines). Solid lines show the true function. The estimation bias for the spatial effects in the outcome equation is illustrated for both methods in the bottom graphs of Figure 3.1. For the spatial effects in the selection equation no differences were visible. Therefore, the corresponding graphs are omitted.
In Table 3.4, empirical root mean squared errors averaged over the simulation runs for separate univariate regressions and the sample selection model are given, where the

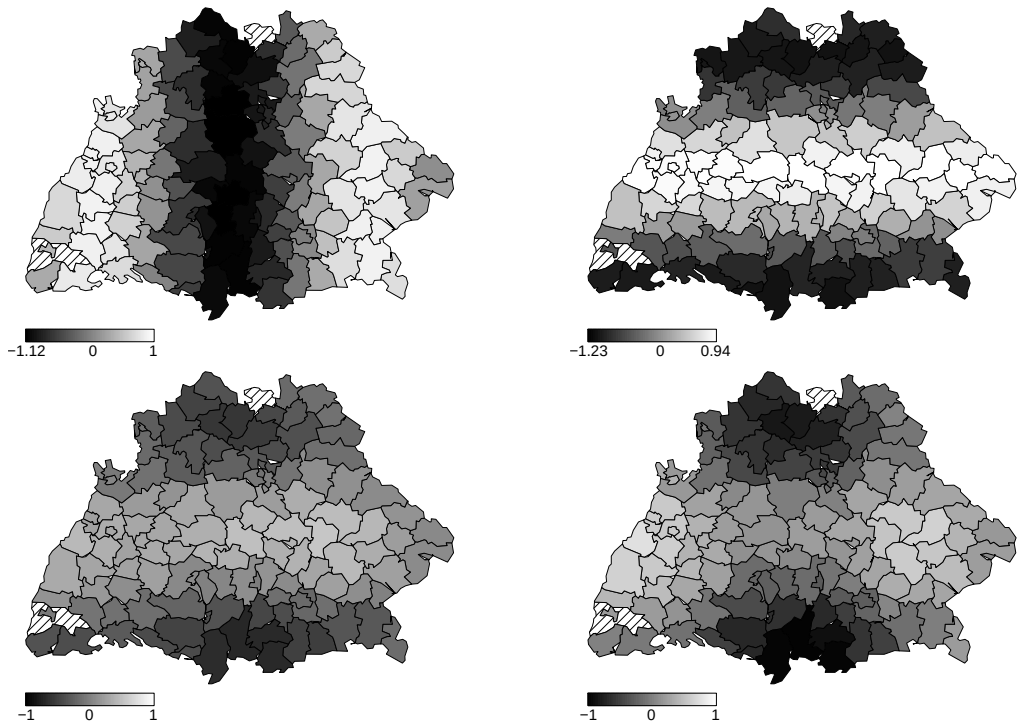

Figure 3.1: Simulation study 2: The first row shows the true spatial effects in the selection equation (left) and the outcome equation (right). In the second row, maps of the estimation bias of the spatial effects in the outcome equation for the sample selection model (left) and univariate regression (right) are shown. In shaded regions, no data were simulated.

empirical root mean squared error for estimates $\hat{f}_r$ from simulation run r is defined as

$$RMSE(\hat{f}_r) = \sqrt{\frac{1}{200}\sum_{i=1}^{200}(\hat{f}_r(x_i) - f(x_i))^2}.$$

Note that the RMSE for the nonparametric functions is based on estimates for 200 fixed covariate values which is necessary due to different missing values of $y_2 = (y_{12}, \ldots, y_{n_2 2})^t$ and consequently of the covariate values. The RMSE of the spatial function $f_{2,\text{spat}}$ is based on estimates for all regions including those missing in observations available for the outcome equation. For missing spatial regions, estimates are obtained by sampling from the corresponding full conditional, i.e. by predicting estimates also for these regions. Thus, uncertainty in the estimate for the complete spatial function is adequately reflected

and results are comparable between simulation runs with different missing regions.
The following conclusions can be drawn:

- For functions f_{21} and $f_{2,\text{spat}}$ which enter the outcome equation with the same covariates as in the selection equation, estimates are severely biased and the mean squared errors are high when using univariate regression. More precisely, when using univariate regression strong true spatial effects $f_{2,\text{spat}}$ are not recovered and the magnitude of the effects is underestimated. This is particularly the case in regions that are likely to be unobserved and that have negative effects in the outcome equation as well as in regions where censoring is less likely and that have positive effects in the outcome equation. The sample selection model considerably reduces the estimation bias and the RMSE for both functions.

- For the remaining functions, no clear differences between the fits and mean squared errors obtained by univariate regression and the sample selection model can be observed, although the sample selection model yields minimally better results.

Also the average coverage rates of pointwise credible intervals based on nominal levels of 80% and 95% were calculated. For the biased fits of functions f_{21} and $f_{2,\text{spat}}$ in univariate regression, the coverage rates were clearly below the nominal level, while those in the sample selection model were above the nominal level. For the other functions, the coverage rates of both methods were virtually equal and except for function f_{22} above the nominal level.
Summing up, compared to separate univariate regressions, the sample selection model reduces the estimation bias and the mean squared error for effects of covariates that are included in both equations and leads to reliable uncertainty estimates.

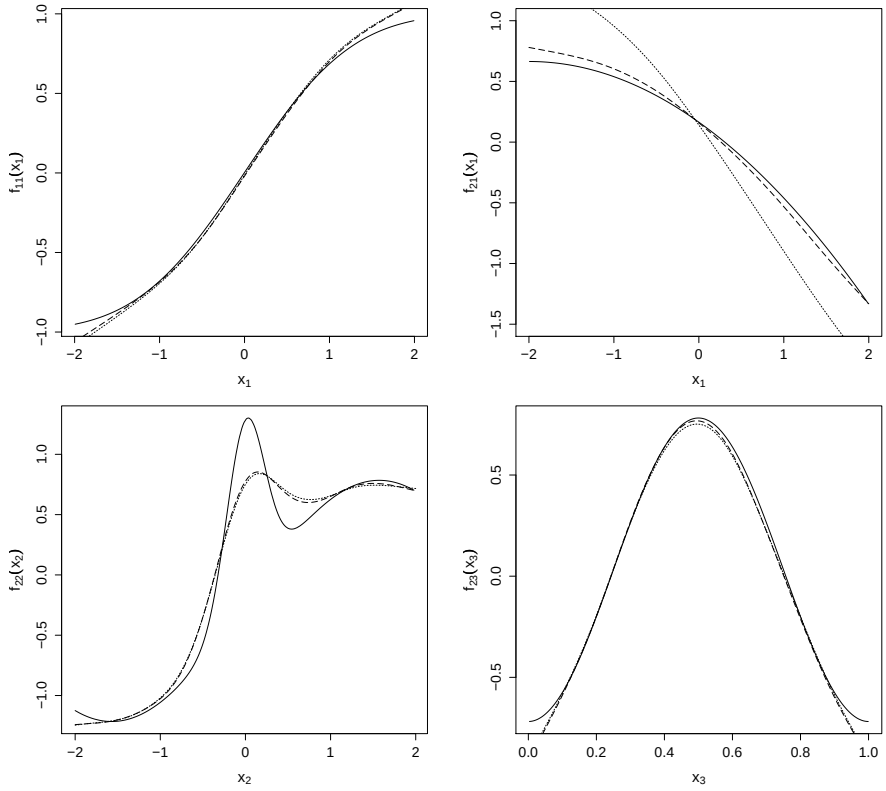

Figure 3.2: Simulation study 2: Averaged fits of the nonparametric functions. The dashed lines show averaged posterior mean estimates of the Bayesian sample selection model and the dotted lines show those obtained by Bayesian univariate regression. The solid lines are the true functions. Note that the curves only differ minimally in some cases which is why the lines overlay and the curves can hardly be distinguished.

	Selection Equation		Outcome Equation			
	f_{11}	$f_{1,\text{spat}}$	f_{21}	f_{22}	f_{23}	$f_{2,\text{spat}}$
Univ.	0.1308	0.4053	0.4491	0.2504	0.1456	0.5164
SSM	0.1224	0.3986	0.1556	0.2421	0.1432	0.4166

Table 3.4: Simulation study 2: Empirical root mean squared errors for univariate regressions (Univ.) and the sample selection model (SSM).

3.5 Relief Supply in Earthquake-Affected Communities in Pakistan

3.5.1 Data Sources and Data Preparation

On October 8, 2005, Pakistan was hit by a magnitude 7.6 earthquake centered in Azad Jammu Kashmir (AJK) province. The earthquake killed at least 73,000 people and made millions homeless. A large-scale internationally coordinated response followed to provide the affected communities with relief supply. Approximately 90 distribution agencies asked the United Nations Joint Logistics Center (UNJLC) to coordinate the movements of their cargo while other agencies coordinated the response independently. The data set considered in the following contains information only on the deliveries coordinated by the UNJLC. This restriction, for example, implies that larger settlements are underrepresented in the data since these were mainly accommodated by providers not coordinated by the UNJLC (particulary the Pakistani armed forces). Between 28 October 2005 and 18 May 2006, the UNJLC coordinated deliveries of goods from 32 origins to 219 destinations within 87 Union Councils in the operation zones Batagram, Mansehra, Muzaffarabad and Bagh. The October observations were considered incomplete and have therefore been removed from the data set, resulting in observations for a time span of 199 days.

The deliveries are divided into commodity types such as food, kitchen supplies and water (commodity type 1) or tools, shelter and clothing (commodity type 2). In the following, we will consider the quantities delivered for each of the two commodity types as response variables in the outcome equations of two separate models (844 and 430 observations, respectively) although it would in principle be possible to combine both commodity types into one joint model as outlined in Section 3.6. Both responses are measured in metric tons and were transformed logarithmically to match the assumption of a Gaussian distribution. The dependent variables in the selection equations represent the decision to deliver the considered commodity type on a given day. To be more specific, the binary selection indicator equals one, if in a certain region on a certain day a movement of the respective commodity type took place, otherwise it is coded as zero. In summary, we obtain $87 \cdot 199 = 17,313$ observations for the 87 Union Councils and 199 days constituting the observation period, leading to degrees of censoring of approximatively 95% and 97.5%, respectively.

Covariates from external sources were added to the UNJLC database for the analyzes. The covariates can be grouped into needs-related and logistics-related variables. Since no immediate measures for survivor needs are available, estimated pre-disaster population

size is employed as a proxy variable for the size of the affected population in a Union Council. In addition, the modified Mercalli index (MMI) measuring seismic strength is considered a proxy of vulnerability. Rugosity is included as a proxy for poverty since mountain villages and dispersed-homestead communities are assumed to be poorer than valley-floor communities. Logistics-related covariates measure the height above sea level, the distance to the responsible supply hub, the available helicopter capacity (in metric tons) and the accessibility of the community by road on a particular day. Note that the latter two covariates change over time although we will suppress time-dependency in the notation. For a more detailed description of the data and a discussion of its implications like the construction of commodity types and the choice of needs and logistical factors see Benini, Conley, Dittemore & Waksman (2006) and Benini, Conley, Dittemore & Waksman (2009).

3.5.2 Model and Prior Settings

For both commodity types, we estimated geoadditive sample selection models. All needs-related variables are included as time-varying effects, distance from supply hub is included nonlinearly and the remaining logistics-related variables enter the model parametrically. The predictor specification is completed by including a spatial effect based on the Union Council, leading to the predictor

$$\begin{aligned}\eta_{ij} &= \gamma_{j0} + \gamma_{j1} height_i + \gamma_{j2} lnheli_i + \gamma_{j3} acc_i + f_j(dist_i) \\ &\quad + pop_i g_{j1}(t) + MMI_i g_{j2}(t) + rug_i g_{j3}(t) + f_{j,\text{spat}}(s_i), \quad j = 1, 2,\end{aligned}$$

where $\gamma_{j0}, \ldots, \gamma_{j3}$ correspond to intercept and parametric effects for elevation above sea level (*height*), logarithm of helicopter capacity (*lnheli*) and road access (*acc*, binary). f_j is the nonparametric effect of distance to the next supply hub (*dist*), g_{j1}, g_{j2}, g_{j3} are the time-varying effects of population size (*pop*), seismic strength (*MMI*) and rugosity (*rug*) and $f_{j,\text{spat}}$ represents the spatial effect. Time-varying effects were assigned to the needs-related variables to determine the temporal variation of the impact of survivor needs (as compared to logistic convenience) within the observation time.
Cubic splines with second order random walk prior and 30 knots were considered for both the nonparametric and the time-varying effects. For the hyperpriors of smoothing and error variances, the prior parameters are fixed at $a = b = 0.001$ and $a_{\sigma_{2|1}} = b_{\sigma_{2|1}} = 0.001$, respectively. For the normal prior of the covariance, we set $m_{\sigma_{21}} = 0$ and $s^2_{\sigma_{12}} = 10$. After a burn-in period of $20,000$ iterations, 80,000 additional iterations were conducted, recording only every 80th iteration to reduce autocorrelations. Inferences are therefore based on 1,000 samples considered to be approximately independent.

	Food, Kitchen Supplies & Water			Construction Material & Tools		
	Estimate	Std.Dev.	p-value	Estimate	Std.Dev.	p-value
Selection equation						
(Intercept)	−10.0451	2.1097	0.000	−8.7294	2.6672	0.000
height	0.0014	0.0005	0.000	0.0001	0.0005	0.870
lnheli	0.7169	0.2909	0.022	0.2857	0.3559	0.460
acc	0.0759	0.0707	0.302	0.2023	0.0731	0.006
Outcome equation						
(Intercept)	35.3435	6.5789	0.000	31.2483	9.1100	0.000
height	−0.0004	0.0006	0.492	−0.0005	0.0005	0.276
lnheli	−1.1028	0.9751	0.290	−1.4555	1.3672	0.252
acc	−0.1831	0.1751	0.286	0.0749	0.1820	0.642
Correlation	-0.9105	0.0299		-0.8662	0.0851	

Table 3.5: Parametric estimates with standard deviations and two-sided Bayesian p-values.

3.5.3 Results

Parametric estimates for both commodity types are summarized in Table 3.5. Graphs of the estimated nonparametric effects are given in Figures 3.3 and 3.4. Maps of UNJLC operation zones Batagram, Mansehra, Muzaffarabad and Bagh with estimated spatial effects are given in Figures 3.5 and 3.6 where the top graphs show the posterior mean estimates of Union Council-specific regional effects and the bottom graphs show maps of significance based on nominal levels of 80%. To obtain the latter from the sampled parameters, 80% credible intervals based on the corresponding sample quantiles were derived. If the credible interval was strictly positive, this is coded as +1 whereas strictly negative intervals are coded as −1. Intervals containing zero are coded as 0. Consequently, regions with nonsignificant regional effects are colored in grey, those with negatively significant effects are colored in black and those with positively significant effects are colored in white. Note that the maps show a larger part of Pakistan to ease the localization of the earthquake-affected regions. Shaded Union Councils have not been used in the estimation process.

In both models, a correlation of the errors of about −0.9 indicates the presence of selection bias and a strong influence of the delivery probability on the amount delivered. In other words, it is suggested that communities with rarer deliveries were compensated with larger amounts in each delivery or, vice versa, that frequent deliveries came along with lower amounts in each delivery. An alternative explanation might be that smaller requests were honored more easily while larger requests had to be rejected more

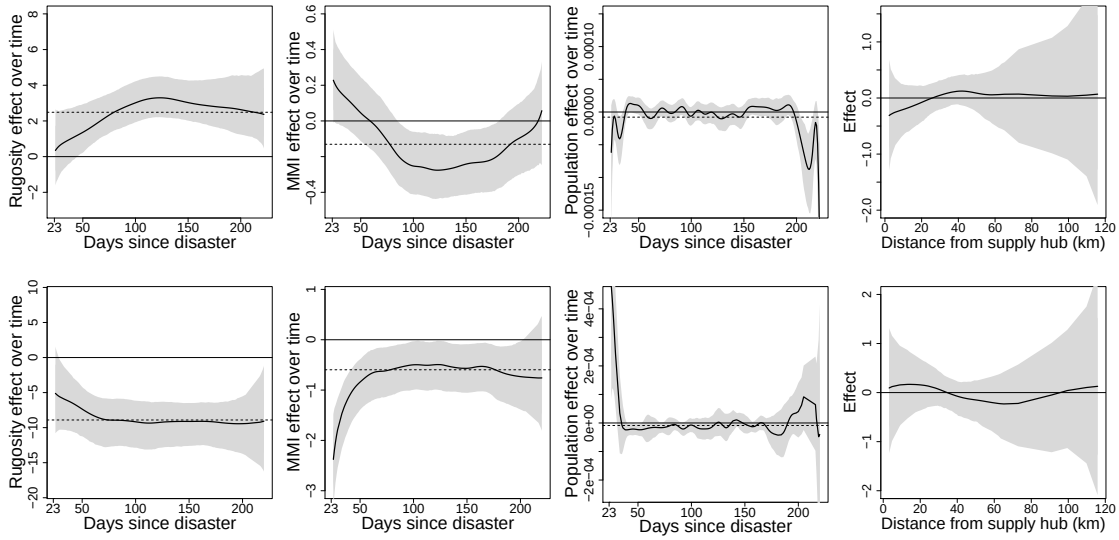

Figure 3.3: Food, kitchen supplies & water: Estimated nonparametric effects in the selection equation (top graphs) and outcome equation (bottom graphs). The right column shows the effect of the logistics-related variable distance and the remaining show time-varying effects of the needs-related variables. Shown are posterior means with 95% pointwise credible intervals. The dotted lines show the mean levels of the functions.

frequently.

Regarding the logistics-related covariates with linear effects, helicopter capacity has a positive effect with posterior probability larger than 95% on the decision to deliver food, kitchen supplies and water. Road access has a positive effect with posterior probability larger than 95% on the decision to deliver construction material and tools. This may reflect a preference to carry construction material by trucks and food by helicopters. Base elevation has a positive effect with posterior probability larger than 95% on the decision to deliver which might also capture the consideration of expected poverty. While all coefficients are positive in the selection equation, their counterparts in the outcome equations are mostly negative. This might imply that a high number of deliveries comes along with less weight in every delivery which coincides with the interpretation of the correlation between the errors. The nonparametric effect of the distance from the responsible supply hubs does not obey a clear structure. In commodity type construction material and tools, there might be an indication of a positive effect of distance on both

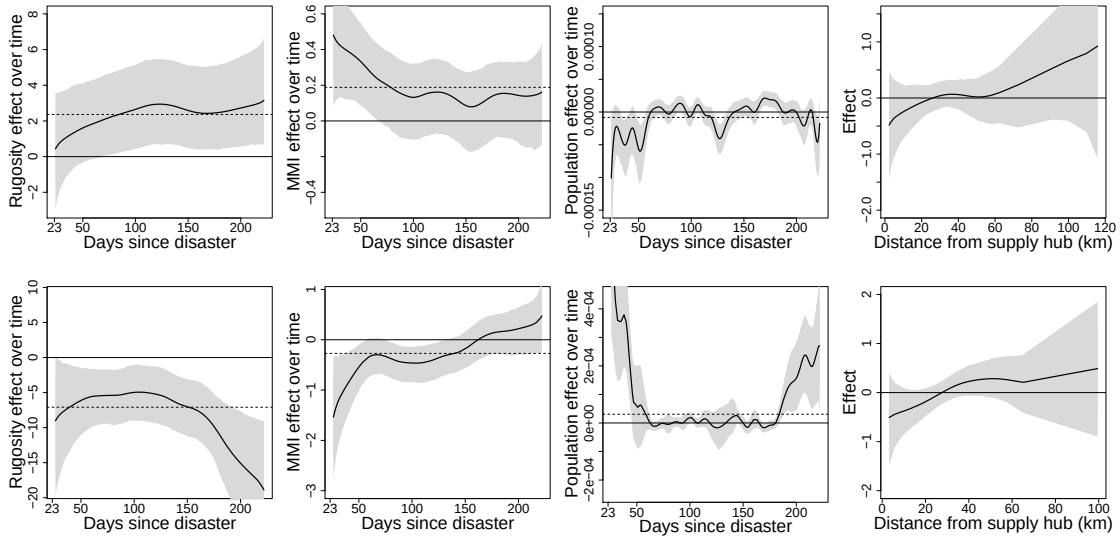

Figure 3.4: Construction material & tools: Estimated nonparametric effects. Graphs are arranged as in Figure 3.3.

response variables, however with wide pointwise credible intervals overlapping zero.
We now turn to the needs-related variables whose effects are assumed to vary over time. While no clear effect of population can be observed for commodity type food, kitchen supplies and water, the graph for commodity type construction material and tools shows positive effects on the amount delivered at the beginning and end of the time period. This suggests that Union Councils with a large population received bigger deliveries than Union Councils with a smaller population at the beginning and end of the time period. Regarding the effects of rugosity and MMI on the decision to deliver relief, the same pattern described by Benini, Conley, Dittemore & Waksman (2009) can be observed: Initially, Union Councils close to the epicenter of the earthquake obtained priority by the agencies, but approaching winter, the influence of rugosity (proxying poverty) increased. Regarding the effects of these factors on the amount of delivered supply, this pattern appears to be reversed, but is associated with high uncertainty. The mean levels of rugosity and MMI are negative. Together with their effect on the decision to ship, this might suggest that poor settlements and Union Councils close to the epicenter received more frequent but smaller deliveries.
The maps of the spatial effects in both equations and models show positive effects in Union Councils close to the epicenter. This might suggest that the MMI does not fully

explain the influence of local damage. The black-colored Union Councils in the very east of the Azad Jammu Kashmir region on the maps for construction material and food on the decision to ship did not receive deliveries at all.

3.5.4 Discussion

One of the major aims in delivering relief supply will be that the delivered goods reach the people in need while logistical factors such as capacity restrictions impose natural restrictions. Our results indicate that in fact not only logistical but also needs-related factors seem to have been taken into account particularly when deciding whether to deliver. Intuitive interpretations of the results are possible in most cases. However, some questions arise concerning the validity of the model and the data used.
The first issue is the construction of the dependent variable in the selection equation. For both commodity types, the number of censored observations is very high, contrasting $16,469$ censored with 844 uncensored observations and $16,913$ censored with 430 uncensored observations for food, kitchen and water supply and construction material and tools, respectively. The high amount of censoring is induced by the construction of the decision indicator where it is assumed that on every day in every Union Council there is a need (and therefore an implicit request) for relief supply. Of course it would be more realistic to work with actual delivery requests but these are not recorded in the data and can hardly be imagined to be collected in a natural disaster area as post-earthquake Pakistan. Moreover, we expect the construction of decision indicators to be related mostly to a shift in the intercept of the selection equation while the covariate effects should be less affected. A second problem with the preparation of the data is that deliveries taking more than one day are counted as observations on each day during the delivery. This might induce bias in the estimates due to observations in Union Councils that are far from the supply hubs and might in particular impact the coefficient associated with the variable distance.
Several of the explanatory variables considered in our models are only proxies for the covariates of interest. For example, rugosity is considered to approximate poverty while the modified Mercalli index proxies vulnerability. While the use of proxy variables leaves some doubts about the estimated effects for covariates such as poverty and vulnerability, the inclusion of the spatial effect actually allows to cover some of the associated uncertainty. For example, we found that the spatial effects hint at both increased probabilities of positive delivery decisions and higher delivery amounts close to the epicenter. This might be related to the fact that the modified Mercalli index fails to capture the full picture of local damage.

Uncertainty about the general validity of our results arises also from the fact that our data set only contains data from agencies coordinated by the UNJLC. In particular, this led to an underrepresentation of larger cities that were mostly accommodated by the Pakistani armed forces. We therefore reestimated the geoadditive sample selection models excluding cities with a population larger than $100,000$. Parametric estimates partly differed (in their magnitude but not their sign) while time-varying effects showed a shift of the overall level but not of the general functional form. In summary, there are no dramatic changes in the interpretation of the estimation results, despite the differences in numerical values.

A final issue is concerned with a general phenomenon in Bayesian sample selection models: Autocorrelations of parameters sampled in the MCMC algorithm typically do not disappear even with large numbers of iterations in particular for the estimates in the selection equation and the components of the covariance matrix and when the correlation between outcome and selection equation is high. We have tried to alleviate the problem by considering quite long simulation runs and considerable thinning but still uncertainty estimates might be affected by the autocorrelation. Again note that this is a common phenomenon in Bayesian sample selection models and is not induced by the geoadditive structure of the predictors.

Due to these problems, the analysis should be considered exploratory. However, the results are intuitively interpretable and the analysis is an interesting example of the application of the geoadditive sample selection model.

3.6 Outlook & Extensions

We have developed a Bayesian geoadditive sample selection model that allows us to analyze sample selection models with considerable flexibility in setting up the model equation. Based on the same types of prior distributions as considered in this chapter, extensions to surface estimation or the inclusion of random effects could be considered along the lines of structured additive regression as suggested in Fahrmeir, Kneib & Lang (2004). For example, temporal correlations could easily be dealt with by including i.i.d. random effects for the Union Councils if a conditionally Gaussian random effects distribution is chosen. In that case, by assigning an inverse Wishart hyperprior to their variance, also correlations between the random effects of the two equations could be accounted for. However, we refrained from this in our application because of the high degree of censoring and the resulting small number of observations available in the outcome equation.

Another extension, also dealing with the issue of modeling temporal correlations more

explicitly, would be the inclusion of an AR-type component for the error terms. However, since the error is actually bivariate, one would also have to include cross-correlations leading to a large number of correlation parameters that would only be weakly identified by the data. Still, this issue might deserve further attention and could be a subject of future research.

Due to the latent Gaussian formulation, the sample selection model could also be extended to contain more than two equations. However, with a rising number of equations the number of covariance coefficients gets large such that updating an inverse Wishart type prior easily becomes numerically unstable. As a consequence, the construction of an MCMC sampler that mixes well despite the large number of weakly identified correlation parameters would be a challenge. The latent Gaussian representation could also be used to allow for binary or categorical responses in the outcome equation along the lines of Albert & Chib (1993).

The suggested approach has been implemented in an R package, see Section 5.2.2 for details.

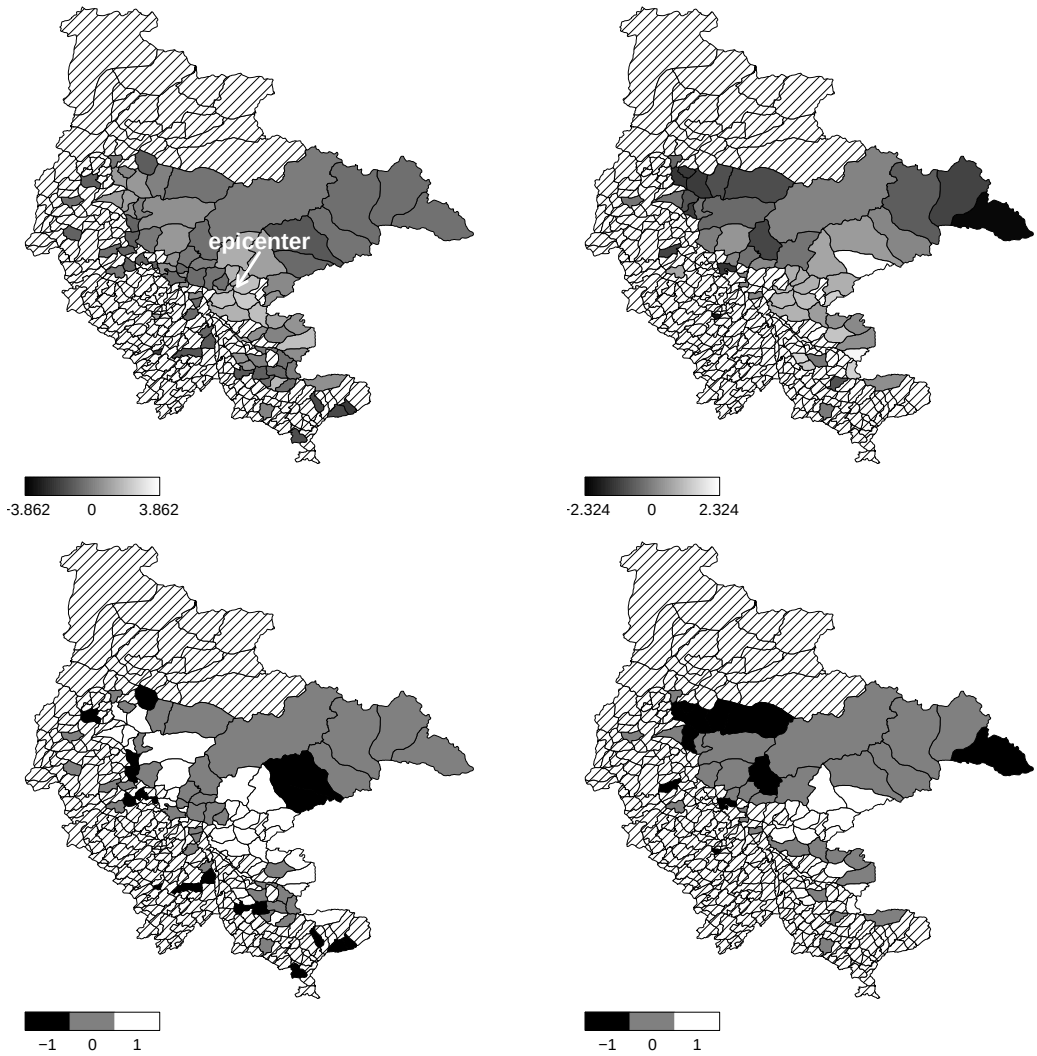

Figure 3.5: Food, kitchen supplies & water: Estimated spatial effects in the selection equation (left column) and outcome equation (right column). The top graphs show posterior means and the bottom graphs show maps of significance based on nominal levels of 80%. The arrow in the top left graph points at the approximative location of the epicenter. In shaded regions no observations were made. Thus, they are excluded from the analysis.

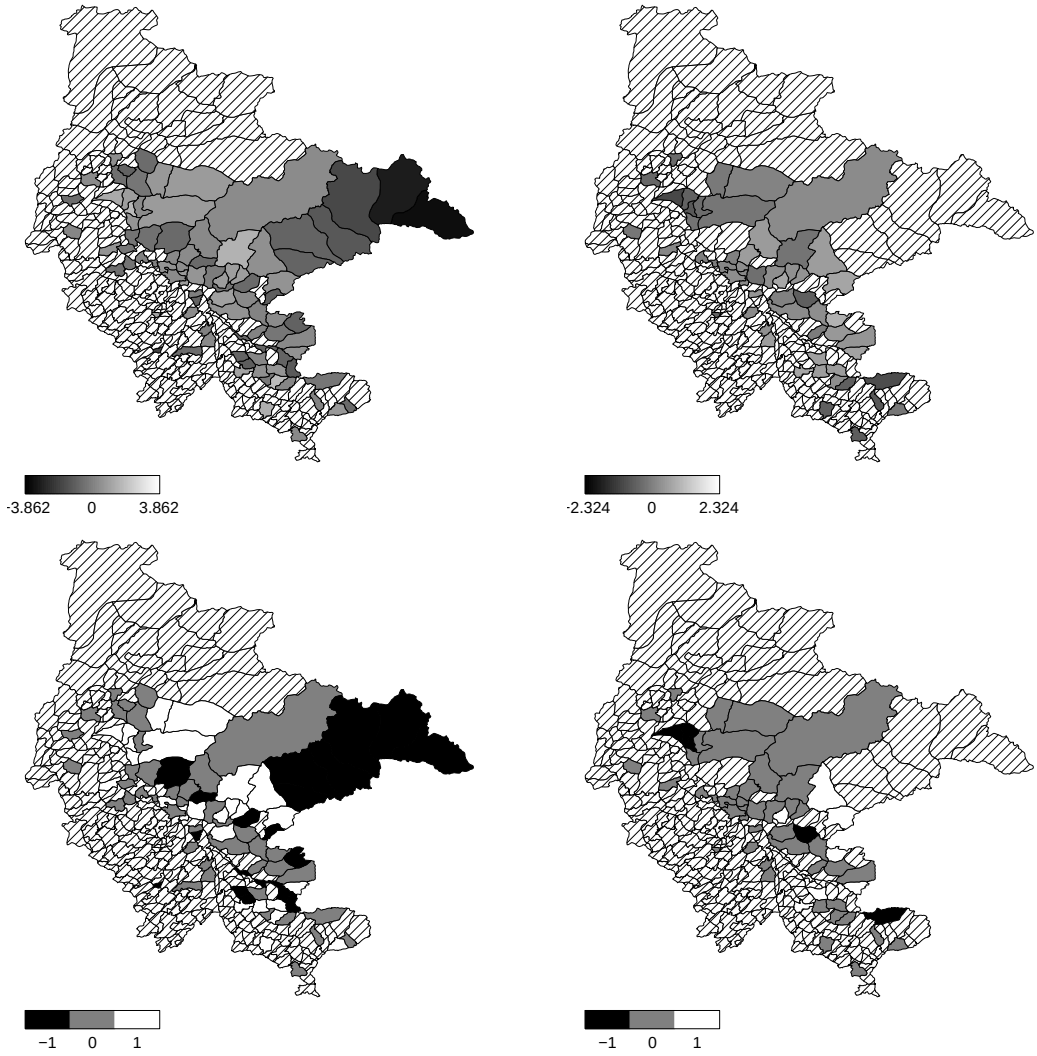

Figure 3.6: Construction material & tools: Estimated spatial effects. Graphs are arranged as in Figure 3.5.

4 Bayesian Nonparametric Instrumental Variable Regression based on Penalized Splines and Dirichlet Process Mixtures

Abstract: *We propose a Bayesian nonparametric instrumental variable approach that allows us to correct for endogeneity bias in regression models where the covariate effects enter with unknown functional form. Bias correction relies on a simultaneous equations specification with flexible modeling of the joint error distribution implemented via a Dirichlet process mixture prior. Both the structural and instrumental variable equation are specified in terms of additive predictors comprising penalized splines for nonlinear effects of continuous covariates. Inference is fully Bayesian, employing efficient Markov Chain Monte Carlo simulation techniques. The resulting posterior samples do not only provide us with point estimates, but allow us to construct simultaneous credible bands for the nonparametric effects, including data-driven smoothing parameter selection. In addition, improved robustness properties are achieved due to the flexible error distribution specification. Both these features are extremely challenging in the classical framework, making the Bayesian one advantageous. In simulations, we investigate small sample properties and an investigation of the effect of class size on student performance in Israel provides an illustration of the proposed approach which is implemented in an R package bayesIV.*

4.1 Introduction

One of the most frequently encountered problems in regression analysis in particular in case of observational data common in the social sciences are endogenous regressors, i.e. explanatory variables that are correlated with the unobservable error term. Sources of this correlation include omitted variables that are associated with both regressors and response (confounder), measurement error, reverse causality and sample selection.

Neglecting the resulting asymptotically not vanishing endogeneity bias by using standard regression techniques can lead to severely misleading inference. In the parametric regression context, the omnipresence of this problem has led to a vast corresponding literature. Two-stage least squares (2SLS) and generalized methods of moments (GMM) estimators in combination with instrumental variables, i.e. additional covariates that are uncorrelated to the error term but reasonably strongly associated to the endogenous covariate, are then routinely applied (see e.g. Wooldridge, 2002). These approaches do not necessarily make distributional assumptions for the error term (for point estimation) but intrinsically rely on linearity of all effects, which is frequently not justified by subject-matter considerations (see also Kleibergen & Zivot (2003) for an overview over Bayesian parametric methods and their association to the related frequentist methods). Thus, in recent years an increasing number of approaches to nonparametric instrumental variable regression has appeared, see Blundell & Powell (2003) for an excellent survey and also Horowitz (2011) including a discussion on implications on inference in misspecified parametric models making a strong case for nonparametric estimation. However, still these methods are rarely used in practice mainly due to a lack of easily available implementations and the need of user assistance, i.e. they typically strongly depend on tuning parameters that can hardly be estimated automatically. This chapter addresses these issues by providing a Bayesian framework which routinely allows the automatic choice of tuning parameters and the construction of simultaneous credible bands for the quantification of the uncertainty of function estimates. Simultaneous credible bands are the Bayesian analogue to simultaneous confidence bands which are important in order to assess the uncertainty of an entire curve estimate and study the relevance of an effect, for example. Pointwise confidence bands, which are almost exclusively used for this purpose, will understate this uncertainty and can thus lead to erroneous identifications of nonlinear effects.

In general, the available nonparametric frequentist approaches can be split into two groups that are based on different identification restrictions: control function approaches and instrumental variable approaches.

The *control function approach* (Newey, Powell & Vella, 1999, Pinkse, 2000 and Su & Ullah, 2008) is directly related to the simultaneous equations literature. For simplicity, for the remainder of the introduction we consider the model with a single endogenous covariate

$$y_2 = f_2(y_1) + \varepsilon_2, \qquad y_1 = f_1(z_1) + \varepsilon_1 \tag{4.1}$$

with response y_2, covariate y_1 and instrumental variable z_1 with effects of unknown functional form f_2 and f_1, respectively, and random errors ε_2 and ε_1. Endogeneity

bias arises if $\mathrm{E}(\varepsilon_2|\varepsilon_1) \neq 0$. Then, assuming identification restrictions $\mathrm{E}(\varepsilon_1|z_1) = 0$ and $\mathrm{E}(\varepsilon_2|\varepsilon_1, z_1) = \mathrm{E}(\varepsilon_2|\varepsilon_1)$, it follows

$$\mathrm{E}(y_2|y_1, z_1) = f_2(y_1) + \mathrm{E}(\varepsilon_2|y_1, z_1) = f_2(y_1) + \mathrm{E}(\varepsilon_2|\varepsilon_1, z_1) = f_2(y_1) + \mathrm{E}(\varepsilon_2|\varepsilon_1) = f_2(y_1) + v(\varepsilon_1) \tag{4.2}$$

where $v(\varepsilon_1)$ is a function of the unobserved error term in the first equation. This has motivated the following two-stage estimation scheme: In a first step, estimated residuals $\hat{\varepsilon}_1$ are determined from $y_1 - \hat{f}_1(z_1)$ using any nonparametric estimation technique for estimating the nonlinear function $\hat{f}_1(z_1)$. In a second step, an additive model (e.g. Hastie & Tibshirani, 1990) with response variable y_2 is estimated, where in addition to y_1 the estimated residuals $\hat{\varepsilon}_1$ are included as a further covariate. Disadvantages of this two-stage approach include the difficulty to incorporate the uncertainty introduced by estimating the parameters in the first step when constructing confidence bands in the second step. In particular, no approach for simultaneous confidence bands that accounts for this uncertainty has been proposed to date. In addition, automatic smoothing parameter selection for the control function $v(\varepsilon_1)$ is difficult since common selection criteria like cross-validation or plug-in estimators focus on minimizing the error in predicting the response variable y_2 while we are interested in achieving a precise estimate for $v(\varepsilon_1)$ to yield full control for endogeneity. Finally, outliers and extreme observations in ε_1 may severely affect the endogeneity correction and therefore some sort of robustness correction (such as trimming of the residuals) might be necessary (Newey, Powell & Vella, 1999).

A completely different strategy is to assume $\mathrm{E}(\varepsilon_2|z_1) = \mathrm{E}(y_2 - f_2(y_1)|z_1) = 0$ leading to the *instrumental variables approach*, see for example Newey & Powell (2003). Here, an ill-posed inverse problem has to be solved creating the need for an additional regularization parameter. Data-driven simultaneous selection of the smoothing parameter and the regularization parameter is still an open question (Darolles, Fan, Florens & Renault, 2011). Again, also construction of simultaneous confidence bands is difficult, with Horowitz & Lee (2009) being the first attempt. In the remainder of this chapter this approach will not be discussed further.

In the Bayesian framework, most available nonparametric approaches are based on representing the model as simultaneous equations and are thus related to the control function approach. All of these assume bivariate normality of the errors $(\varepsilon_1, \varepsilon_2) \sim \mathrm{N}(0, \Sigma)$ (e.g. Chib & Greenberg, 2007, Chib, Greenberg & Jeliazkov, 2009 and Koop, Poirier & Tobias, 2005). Then, both equations in (4.1) are estimated simultaneously in a Gibbs-sampling scheme, facilitating the estimation of smoothing parameters and credible bands. Thus, the control function is not explicitly estimated but is given implicitly by the conditional

error distribution. However, bivariate normality implies linearity of this conditional expectation since $E(\varepsilon_2|\varepsilon_1) = \frac{\sigma_{12}}{\sigma_1^2}\varepsilon_1$, where $\sigma_{12} = \text{Cov}(\varepsilon_{1i}, \varepsilon_{2i})$ is the covariance of the error terms and $\sigma_1^2 = \text{Var}(\varepsilon_{1i})$. As a consequence, the control function is implicitly restricted to be linear in ε_1, corresponding to the assumption that a hypothetical (unknown) omitted variable inducing the endogeneity bias has a linear effect on the response. This assumption seems to be rather restrictive, in particular when allowing for effects of unknown functional form for all of the observed explanatory variables. Relaxing the distributional assumption (as we will do in the following) results in allowing the omitted variables to have an effect of unknown functional form as well. Note that although 2SLS procedures interpreted in their control function representation in the fully parametric context (where all functions are restricted to be linear and estimation is based on ordinary least squares) do not make assumptions on the marginal distributions of ε_1 and ε_2. However, they still rely on linearity of the conditional expectation $\text{E}(\varepsilon_2|\varepsilon_1)$. Another common source for non-normality of the error terms are outliers and thus robustness issues of methods relying on bivariate normality are a serious concern. As a consequence, Conley, Hansen, McCulloch & Rossi (2008) propose the application of a Dirichlet process mixture (DPM) prior (Escobar & West, 1995) to obtain a flexible error distribution, but they still rely on linear covariate effects.

In this work, we extend their approach by proposing a Bayesian approach based on Markov chain Monte Carlo (MCMC) simulation techniques employing Bayesian P-splines (Lang & Brezger, 2004) for the estimation of flexible covariate effects and a DPM prior for the estimation of a flexible joint error distribution. Univariate regression models with smooth covariate effects and a DPM prior for the error density have been previously considered among others by Chib & Greenberg (2010). Thus, neither we make an assumption on the functional form of the effects (besides a smoothness condition) nor on the distribution of the error terms. Further, we will allow a more flexible choice of prior distributions than Conley, Hansen, McCulloch & Rossi (2008). The Bayesian formulation will enable us to automatically estimate the smoothing parameters in both equations and to construct simultaneous credible bands that do not depend on distributional assumptions. Moreover, through the use of the DPM prior, outliers in the error terms will automatically be downweighted such that improved outlier robustness is provided.

The approach is used to analyze the effect of class size on scholastic achievements of students in Israel following Angrist & Lavy (1999). Thereby, a clearly non-normal bivariate error density warrants nonparametric estimation of the error density in order to ensure proper endogeneity bias correction and valid confidence bands. As already suggested by Horowitz (2011), nonparametric estimation of the relationship in combination

with simultaneous credible bands is important for proper evaluation of the estimation uncertainty and is able to reveal new insights into the relationship.
The remainder of the chapter is organized as follows. In Section 4.2 the considered model is introduced and prior distributions are discussed. Section 4.3 describes Bayesian inference including smoothing parameter determination and construction of simultaneous credible bands. In Section 4.4, small sample properties are explored through simulations and the approach is compared to existing approaches. In Section 4.5, an application to class size effects on student performance is provided and the chapter concludes in Section 4.6.

4.2 Additive Simultaneous Equations Model

We consider an additive simultaneous equations model

$$y_{2i} = \gamma_{20} + f_{21}(y_{1i}) + \sum_{\ell=1}^{q_2} x_{2\ell i}\gamma_{2\ell} + \sum_{\ell=1}^{p_2} f_{2,\ell+1}(z_{2\ell i}) + \varepsilon_{2i} \tag{4.3}$$

$$y_{1i} = \gamma_{10} + \sum_{\ell=1}^{q_1} x_{1\ell i}\gamma_{2\ell} + \sum_{\ell=1}^{q_2} x_{2\ell i}\gamma_{1,q_1+\ell} + \sum_{\ell=1}^{p_1} f_{1\ell}(z_{1\ell i}) + \sum_{\ell=1}^{p_2} f_{1,p_1+\ell}(z_{2\ell i}) + \varepsilon_{1i},\ i = 1,.. \tag{4.4}$$

where y_2 denotes the outcome of primary interest affected by one continuous endogenous variable y_1, q_2 exogenous variables $x_{2\ell}$, $\ell = 1, \ldots, q_2$ with linear effects (typically categorical covariates in dummy or effect coding), and p_2 exogenous continuous covariates $z_{2\ell}$, $\ell = 1, \ldots, p_2$. Both the effect of the endogenous variable y_1 and the effects of the continuous covariates $z_{2\ell}$ are allowed to be of unknown, nonlinear form represented by smooth functions $f_{21}(y_1)$ for the endogenous variables and $f_\ell(z_{2,\ell+1})$, $\ell = 1, \ldots, p_2$ for the exogenous covariates. The same model structure applies to the endogenous variable which is related to parametric effects of covariates $x_{1\ell}$, $\ell = 1, \ldots, q_1$ and $x_{2\ell}$, $\ell = 1, \ldots, q_2$ as well as potentially nonlinear effects of continuous covariates $z_{1\ell}$, $\ell = 1, \ldots, p_1$ and $z_{2\ell}$, $\ell = 1, \ldots, p_2$. To ensure identifiability of the additive model structure, all functions $f_{r\ell}(\cdot)$ are centered around zero.
Endogeneity bias in function $f_{21}(y_1)$ arises when the residuals ε_1 and ε_2 are not independent and the outcome equation is estimated without taking the model for the endogenous variable into account. In the simultaneous equations model, identification relies on the instrumental variables $x_{11}, \ldots, x_{1q_1}$ and $z_{11}, \ldots, z_{1p_1}$ (with the same identification restrictions as in the control function approach). While a bivariate normal distribution for the error terms $(\varepsilon_{1i}, \varepsilon_{2i})$ is a convenient model that enables the inclusion of correlated errors (see for example Chib & Greenberg (2007), Chib, Greenberg & Jeliazkov (2009)

or Koop, Poirier & Tobias (2005)) it implies strong implicit assumptions on the control function as discussed in the introduction. We therefore follow Conley, Hansen, McCulloch & Rossi (2008) and employ a Dirichlet process mixture prior (Escobar & West, 1995) for the joint error distribution which basically allows to specify a hyperprior on the space of potential error distributions. More specifically, our prior choices will yield an infinite mixture of bivariate normals prior for the joint error distribution which is able to virtually cater any (continuous) distribution.
In the following, we discuss prior choices for the parameters involved in our additive simultaneous equations model in more detail.

4.2.1 Parametric Effects

For parametric effects $\gamma_{r\ell}$, $r = 1, 2$, $\ell = 0, \ldots, q_r$, we use diffuse priors $p(\gamma_{r\ell}) \propto \text{const}$ in case of complete lack of prior knowledge. Note that there is abundant literature showing that flat priors in combination with very weak (or even superfluous) instrumental variables (i.e. instruments are not or only very weakly related to y_1) can lead to identification problems (see Chao & Phillips, 1998, Hoogerheide, Kaashoek & Van Dijk, 2007, Kleibergen & Van Dijk, 1998 and Kleibergen & Zivot, 2003) and the use of Jeffrey's prior is then recommended. However, when using Dirichlet process mixtures for the joint error distribution, Jeffrey's prior does no longer take the well known form proportional to the determinant of the cross-product of the design matrix that arises in case of normal error terms. Therefore, we will restrict our analyses to flat priors and recommend to check the explanatory power of instrumental variables in advance (similar as in the frequentist framework). Note, however, that the simulations conducted in Section 4.4 indicate that our simultaneous equations approach works well even in the case of quite weak instruments confirming simulations results of Conley, Hansen, McCulloch & Rossi (2008).
Note that inclusion of random effects for clustered or panel data is straight-forward using normal priors (with zero mean and conjugate prior on the variance parameter).

4.2.2 Nonparametric Effects

Since their introduction by Eilers & Marx (1996), penalized splines have become increasingly popular for representing effects of continuous covariates with unknown, nonlinear form but with a global smoothness assumption on differentiability. While the original motivation was mainly based on computational convenience, the properties of penalized splines have now been thoroughly investigated and are well understood, see for example Kauermann, Krivobokova & Fahrmeir (2009), Reiss & Ogden (2009) and Claeskens,

Krivobokova & Opsomer (2009). We will consider the Bayesian analogue to penalized splines as introduced by Lang & Brezger (2004). Therefore we assume that each of the smooth functions $f_{r\ell}(x)$ of some covariate $x \in \{y_1, z_{11}, \ldots, z_{1p_1}, z_{21}, \ldots, z_{2p_2}\}$ can be represented by a suitable spline function, i.e. $f_{r\ell}(x) \in S(d_{r\ell}, \kappa_{r\ell})$, where $S(d_{r\ell}, \kappa_{r\ell})$ denotes the space of spline functions of degree $d_{r\ell}$ with knots $\kappa_{r\ell} = \{x_{\min} < \kappa_1 < \kappa_2 < \ldots < \kappa_{K_{r\ell}} < x_{\max}\}$. Since $S(d_{r\ell}, \kappa_{r\ell})$ is a $(K_{r\ell} + d_{r\ell} + 1)$-dimensional vector space (a subspace of all $d_{r\ell}$-times continuously differentiable functions), $f_{r\ell}(x)$ can then be represented as a linear combination of suitable basis functions $B_k(x)$, i.e.

$$f(x) = \sum_{k=1}^{K_{r\ell}+d_{r\ell}+1} \beta_{r\ell k} B_k(x) = X_{r\ell}\beta_{r\ell}.$$

Due to their simplicity and numerical stability, we will utilize B-spline basis functions in the following.

Although the global smoothness properties are determined by the degree of the spline basis $d_{r\ell}$, the variability of the resulting estimates heavily depends on the location and number of knots. Instead of directly aiming at optimizing the number and position of the knots in a data-driven manner, the penalized spline approach relies on using a generous number of equidistant knots (with the common rule of thumb $K_{r\ell} = \min(n/4, 40)$) in combination with a penalty that avoids overfitting. In the frequentist framework, Eilers & Marx (1996) proposed to penalize the squared q-th order differences of adjacent basis coefficients, thereby approximating the integrated squared q-th derivative of the spline function. In the Bayesian framework, this corresponds to assigning a random walk prior to the spline coefficients with

$$\beta_{r\ell k} = \beta_{r\ell,k-1} + u_k \quad \text{or} \quad \beta_{r\ell k} = 2\beta_{r\ell,k-1} - \beta_{r\ell,k-2} + u_k$$

for first- and second-order random walks with $u_k \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0, \tau^2_{r\ell})$ and non-informative priors for the initial parameters. In this specification, the random walk variance $\tau^2_{r\ell}$ acts as an inverse smoothing parameter with small values corresponding to heavy smoothing while large values allow for considerable variation in the estimated function. In the limiting case of $\tau^2_{r\ell} \to 0$, the estimated function approaches a constant or a linear effect for first and second order random walk priors, respectively. From the random walk specification, the joint prior distribution for the coefficient vector $\beta_{r\ell}$ can be derived as a partially

improper multivariate Gaussian distribution with density

$$p(\beta_{r\ell}|\tau^2_{r\ell}) \propto \left(\frac{1}{2\tau^2_{r\ell}}\right)^{\frac{\text{rank}(\Delta_{r\ell})}{2}} \exp\left(-\frac{1}{2\tau^2_{r\ell}}\beta^t_{r\ell}\Delta_{r\ell}\beta_{r\ell}\right)$$

where $\Delta_{r\ell}$ is the penalty matrix given by the cross-product of a difference matrix $D_{r\ell}$ of appropriate order, i.e. $\Delta_{r\ell} = D^t_{r\ell}D_{r\ell}$.
To complete the fully Bayesian prior specification, a prior on $\tau^2_{r\ell}$ has to be assigned to include estimation of the smoothing variance and therefore to allow for a data-driven amount of smoothness. We choose a conjugate inverse-gamma distribution with shape and scale parameters $a_{\tau_{r\ell}}$ and $b_{\tau_{r\ell}}$, i.e. $\tau^2_{r\ell} \sim \text{IG}(a_{\tau_{r\ell}}, b_{\tau_{r\ell}})$, and will discuss the choice of smoothing parameters in more detail in Section 4.3.3.

4.2.3 Joint Error Distribution

The standard approach in the Bayesian nonparametric simultaneous equations literature for modeling the joint error distribution of $(\varepsilon_{1i}, \varepsilon_{2i})$ is to assume bivariate normal errors $(\varepsilon_{1i}, \varepsilon_{2i}) \sim \text{N}(0, \Sigma),\ i = 1, \ldots, n$ with constant covariance matrix Σ which is assumed to be a priori inverse-Wishart distributed $\Sigma \sim \text{IW}(s_\Sigma, S_\Sigma)$ where IW denotes the inverted-Wishart distribution parameterized such that (for the bivariate case) $\text{E}(\Sigma) = S_\Sigma^{-1}/(s_\Sigma - 3)$.
As mentioned in the introduction, assuming bivariate normality induces strong implicit assumptions on the control function and a violation of these assumptions can have severe impact on the general results and in particular the endogeneity correction. An obvious first relaxation is to use a finite mixture of K^{**} Gaussian components with mixing proportions $\pi_1, \ldots, \pi_{K^{**}}$ and component-specific (nonconstant) means and covariances $\mu_c,\ \Sigma_c,\ c = 1, \ldots, K^{**}$:

$$(\varepsilon_{1i}, \varepsilon_{2i})|\pi_1, \mu_1, \Sigma_1, \ldots, \pi_{K^{**}}, \mu_{K^{**}}, \Sigma_{K^{**}} \text{ i.i.d. } \sum_{c=1}^{K^{**}} \pi_c \,\text{N}(\mu_c, \Sigma_c), \quad \sum_{c=1}^{K^{**}} \pi_c = 1.$$

Though being already quite flexible, this model introduces the problem of selecting the number of mixture components K^{**}. In addition, the number of components is assumed to be fixed as $n \to \infty$ which is an undesired property in the given setting. To remedy both issues, we consider a Gaussian Dirichlet Process Mixture (Escobar & West, 1995) which can be interpreted as the limiting case of a finite mixture model as $K^{**} \to \infty$ (Neal, 2000). More specifically, we assume an infinite mixture model with the following

hierarchy:

$$\begin{aligned}
(\varepsilon_{1i}, \varepsilon_{2i}) \quad & \text{i.i.d.} \quad \sum_{c=1}^{\infty} \pi_c \, \mathrm{N}(\mu_c, \Sigma_c) \\
(\mu_c, \Sigma_c) \quad & \text{i.i.d.} \quad G_0 = \mathrm{N}(\mu | \mu_0, \tau_\Sigma^{-1}\Sigma) \, \mathrm{IW}(\Sigma | s_\Sigma, S_\Sigma) \\
\pi_c \quad & = \quad v_c \left(1 - \sum_{j=1}^{c-1} (1 - \pi_j) \right) = v_c \prod_{j=1}^{c-1} (1 - v_j), \quad c = 1, 2, \ldots \\
v_c \quad & \text{i.i.d.} \quad \mathrm{Be}(1, \alpha).
\end{aligned}$$

In this specification, the mixture components are assumed to be i.i.d. draws from the base measure G_0 (given by a normal-inverse Wishart distribution) of the Dirichlet process (DP) while the mixture weights are generated in a stick-breaking manner based on a Beta distribution depending on the concentration parameter $\alpha > 0$ of the Dirichlet process. The concentration parameter α determines the strength of belief in the base distribution G_0, which is the expectation of the Dirichlet process around which more mass will be concentrated for large α since the variance of the Dirichlet process decreases with α.
In order to emphasize the capability of the prior to model means and covariances varying with observations, we can also express the implied hierarchy by $(\varepsilon_{1i}, \varepsilon_{2i})|(\mu_i, \Sigma_i) \sim \mathrm{N}(\mu_i, \Sigma_i),\ i = 1, \ldots, n$, with $(\mu_i, \Sigma_i)|G \overset{i.i.d.}{\sim} G$ and $G \sim \mathrm{DP}(\alpha, G_0)$ with constructive representation $G = \sum_{c=1}^{\infty} \pi_c \delta_{(\mu_c, \Sigma_c)}$ (Sethuraman, 1994), where δ_θ is a unit point mass at θ. Although we are dealing with an infinite mixture, there can be at most n components affiliated with data and therefore most components will in fact be empty and only determined by the prior. More precisely, in a specific data set errors will be clustered together into $K^* \leq n$ clusters with means $\mu_l = (\mu_{1l}, \mu_{2l})^t$ and covariances $\Sigma_l = \begin{pmatrix} \sigma_{1l}^2 & \sigma_{12,l} \\ \sigma_{12,l} & \sigma_{2l}^2 \end{pmatrix}$, $l = 1, \ldots, K^*$. This can be nicely seen by considering the so-called polya-urn scheme (Blackwell & MacQueen, 1973). Let $\theta_1 = (\mu_1, \Sigma_1), \theta_2 = (\mu_2, \Sigma_2), \ldots$ be an (infinite) sequence of i. i. d. draws from G. Then, the predictive distribution of a new θ_{k+1} conditional on the previous values $\theta_1, \ldots, \theta_k$ marginalizing out G is given by

$$\theta_{k+1} | \theta_1, \ldots, \theta_k \sim \frac{\alpha}{\alpha + k} G_0 + \frac{1}{\alpha + k} \sum_{i=1}^{k} \delta_{\theta_i} \tag{4.5}$$

with δ_{θ_i} denoting a unit point mass at θ_i. That is, θ_{k+1} equals to any of the k previous $\theta_1, \ldots, \theta_k$ with probability $\frac{1}{\alpha+k}$ and is drawn from the base distribution G_0 with probability $\frac{\alpha}{\alpha+k}$. Moreover, Equation (4.5) can also be reexpressed in terms of the distribution

of the distinct values known as a so-called Chinese restaurant process. By doing so, it can be shown that a new θ_{k+1} equals to some θ_l with probability $\frac{n_l}{\alpha+k}$ with n_l the number of values already corresponding to θ_l, i.e. the probability is proportional to the cluster size. Besides the clustering property of the Dirichlet process, these probability expressions also demonstrate the important role of the concentration parameter α: The probability to draw a new value and thus the number of distinct components depends on α. Specifically, the expected number of components for a given sample size n is approximatively given by $\mathrm{E}(K^*|\alpha, n) \approx \alpha \log(1 + n/\alpha)$ (Antoniak, 1974). Thus, the concentration parameter α is directly related to the number K^* of unique pairs $(\mu_l,\ \Sigma_l)$ in the data. In order to avoid fixing K^* we therefore estimate α from the data and consequently have to assign a prior on it. The standard conjugate prior for α is a Gamma prior $\alpha \sim \mathrm{Ga}(a_\alpha, b_\alpha)$. Alternatively, a discrete prior on K^* as in Conley, Hansen, McCulloch & Rossi (2008) can be used (which is equally supported by our software). See Conley, Hansen, McCulloch & Rossi (2008) for details.

Since our model includes constants γ_{10} and γ_{20}, we have to ensure that $\mathrm{E}(\varepsilon_{1i}, \varepsilon_{2i}) = 0$ for identifiability. Though centered Dirichlet Process Mixtures could generally be applied for this purpose, we opt to achieve this by choosing $\mu_0 = (0, 0)^t$ and constraining $\sum_{i=1}^n \mu_{1i} = \sum_{i=1}^n \mu_{2i} = 0$. This simple solution allows us to use efficient algorithms for estimation. Note that from an a priori zero mean $\mu_0 = (0, 0)^t$ alone, it does not follow that G has a posterior zero mean. Note also that for incorporation of categorical variables (dummies) in the regression equation, this constraint is equally required. Conley, Hansen, McCulloch & Rossi (2008) avoid the identifiability constraint by omitting the global intercepts, but oversee the unidentifiability of the dummy coefficients in this case. In fact, this fully explains the deviation of their estimated returns to education (Card, 1995) from the 2SLS estimate and replicating their analysis of the relationship between education and wages imposing $\mathrm{E}(\mu_1) = \mathrm{E}(\mu_2) = 0$ results in an estimate barely differing from the 2SLS estimate.

With respect to priors on the parameters in the base distribution G_0, Conley, Hansen, McCulloch & Rossi (2008) propose to choose parameters μ_0, τ_Σ, s_Σ and S_Σ as fixed in order to reduce the computational burden. They argue that by standardizing y_1 and y_2, zero means $\mu_0 = (0, 0)$, a diagonal S_Σ as well as parameters s_Σ and τ_Σ chosen such that components of Σ_c and μ_c may take even extreme values given the data was standardized beforehand, introduce negligible prior information. However, as Escobar & West (1995) emphasize, the prior variance τ_Σ^{-1} (which is closely linked to the bandwidth in kernel density estimation in case of a constant Σ) has a strong impact on the degree of smoothness of the density. For a given number of distinct mixture components in the data (K^*), a small value of τ_Σ allows the means $(\mu_{1l},\ \mu_{2l}),\ l = 1, \ldots, K^*$ to vary more

strongly resulting in a greater chance of multimodality in the error term distribution for fixed Σ_l. Also, τ_Σ may have an effect on the down-weighting of outliers in the conditional mean $\mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i})$ and thus on the influence of outliers on endogeneity bias correction as we will see in Section 4.3.3. In order to express uncertainty about τ_Σ, Escobar & West (1995) therefore propose to choose a conjugate prior $\tau_\Sigma \sim \mathrm{Ga}(a_\Sigma/2, b_\Sigma/2)$.
Finally, the choice of an inverse Wishart prior on S_Σ, $S_\Sigma \sim \mathrm{IW}(s_{S_\Sigma}, S_{S_\Sigma})$, might be desirable.
Our method allows to flexibly choose between fixed and uncertain hyperparameters.

4.2.4 Hyperparameter Choices

From the properties of the inverse Wishart distribution (see e.g. Link & Barker (2005) for a related discussion) it follows that the residual variances (diagonal elements of Σ_l) are a priori inverse gamma distributed, $\sigma_{rl}^2 \sim \mathrm{IG}((s_\Sigma - 1)/2, S_{\Sigma_{rr}}/2)$, $r = 1, 2$ with $S_{\Sigma_{rr}}$ the r-th diagonal element of S_Σ. Further, given S_Σ is diagonal, it follows that the correlation coefficient ρ_l in component l is a priori beta-distributed, $(\rho_l - 1)/2 \sim \mathrm{Be}((s_\Sigma - 1)/2, (s_\Sigma - 1)/2)$. Thus, the prior of the correlation coefficient has a symmetric density around 0 (since the beta distribution parameters are equal) and consequently choosing a diagonal S_Σ results in a zero prior mean for the correlation $\mathrm{E}(\rho_l|\cdot) = 0$. However, the prior distribution of ρ_l also depends on s_Σ. For $s_\Sigma = 3$, we obtain a $\mathrm{Be}(1, 1)$ distribution which is the uniform distribution, for $s_\Sigma < 3$ we obtain a U-shaped distribution and for $s_\Sigma > 3$ a unimodal distribution. Conley, Hansen, McCulloch & Rossi (2008) use as default specification $s_\Sigma = 2.004$ and thus a prior on ρ_l with a U-shaped density. Thus, although in their prior choice errors are uncorrelated in the mean, more probability mass is assigned to correlations close to -1 and 1 than to values close to zero. To avoid such a prior information, we rather choose $s_\Sigma = 3$ such that the prior on ρ_l is uniform over $[-1, 1]$. Alternatively, in certain situations one might want to choose $s_\Sigma > 3$ such that the prior on ρ_l is unimodal and symmetric around zero in order to a priori favor no endogeneity in case of only weak information in the data (and thereby stabilize estimation similar to regularization techniques).
Given $s_\Sigma = 3$ we obtain $\sigma_{rl}^2 \sim \mathrm{IG}(1, S_{\Sigma_{rr}}/2)$ as prior on the residual variances. Taking into account that responses are centered and standardized, we choose diagonal S_Σ, with equal elements such that the inverse Gamma introduces only weak information on the residual variances. In order to choose these elements, we follow Conley, Hansen, McCulloch & Rossi (2008) and choose default $S_{\Sigma_{rr}}$ such that $P(0.25 < \sigma_{rl} < 3.25) = 0.8$ based on the inverse gamma distribution of σ_{rl}^2 keeping in mind that y_1 and y_2 were standardized beforehand. With $s_\Sigma = 3$ we obtain $S_\Sigma = 0.2I_2$ and thus $\sigma_{rl}^2 \sim \mathrm{IG}(1, 0.1)$

as a weakly informative default. Note that with $s_\Sigma = 2.004$ and $S_\Sigma = 0.17 I_2$, Conley, Hansen, McCulloch & Rossi (2008) choose as default a IG$(0.502, 0.085)$-prior on the residual variances. Although imposing an IW-prior on S_Σ instead is conceptually and computationally straight-forward, associated hyperparameter choice is unclear and is therefore not followed in the remainder of the chapter.
Still, specification of τ_Σ remains to be discussed. Given the possible impact of τ_Σ on the smoothness of the density and weighting of outliers, we might want to impose a hyperprior on τ_Σ, $\tau_\Sigma \sim \text{Ga}(a_\Sigma/2, b_\Sigma/2)$. We will follow Escobar & West (1995) and impose a diffuse gamma prior with default hyperparameters $a_\Sigma = 1$ and $b_\Sigma = 100$ which is in contrast to Conley, Hansen, McCulloch & Rossi (2008) who choose a fixed τ_Σ. The impact of estimating τ_Σ versus fixing it will be studied in our simulation study in Section 4.4.1.
With respect to the concentration parameter α, we follow the recommendation of Ishwaran & James (2002) and choose a Gamma prior with hyperparameters $a_\alpha = b_\alpha = 2$ as defaults. This allows both small and large values of α corresponding to many and few mixture components, respectively.
For the smoothing parameters $\tau^2_{r\ell}$ of nonparametric effects we choose the standard noninformative prior $\tau^2_{r\ell} \sim \text{IG}(0.001, 0.001)$ in the following.

4.3 Bayesian Inference

4.3.1 Estimation

Both equations (4.3) and (4.4) can be written in the generic form $y_r = \eta_r + \varepsilon_r$, $r = 1, 2$, with predictors

$$\eta_r = V_r \gamma_r + \sum_{\ell=1}^{\tilde{p}_r} X_{r\ell} \beta_{r\ell}$$

where all parametric effects (including the intercept) in each equation are combined in the design matrix V_r with regression coefficients γ_r whereas the nonparametric effects are represented using B-spline design matrices $X_{r\ell}$ with corresponding basis coefficients $\beta_{r\ell}$ and $\tilde{p}_1 = p_1 + p_2$, $\tilde{p}_2 = p_2 + 1$.
Estimation is carried out by using Gibbs sampling steps in an efficient Markov Chain Monte Carlo implementation. Specifically, given the parameters of the error distribution, full conditionals for the covariate effect parameters in each equation resemble those for the normal heteroscedastic regression model and sampling techniques proposed in Lang & Brezger (2004) (with heteroscedastic errors) can be applied. On the other hand, given the parameter vectors $\beta_{r\ell}, \tau^2_{r\ell}$, $\ell = 1, \ldots, \tilde{p}_r$ and γ_r, $r = 1, 2$, the components of the error

distribution can be obtained using any algorithm for Bayesian nonparametric estimation of bivariate densities based on DPM priors (see Neal (2000) for an overview). Thus, our software allows to choose efficiently implemented algorithms that are called on top of our sampler. More precisely, we use the implementation provided by the R package DPpackage (Jara, Hanson, Quintana, Müller & Rosner, 2011) of two Gibbs sampling algorithms with auxiliary variables given in Neal (2000). In addition, the implementation accompanying Conley, Hansen, McCulloch & Rossi (2008) is integrated.
Full details on all full conditionals are given in the following.

4.3.2 Full Conditionals

In the following, full conditionals for the parameters in the r-th equation, i.e. $r = 1$ for equation (4.4) and $r = 2$ for equation (4.3), are given.

Nonparametric effects The full conditionals for the regression coefficients of the smooth functions are Gaussian

$$\beta_{r\ell}|\cdot \sim \mathrm{N}(\mu_{\beta_{r\ell}}, P^{-1}_{\beta_{r\ell}})$$

with precision matrix

$$P_{\beta_{r\ell}} = X^t_{r\ell}\Sigma^{-1}_{r|-r}X_{r\ell} + \frac{\Delta_{r\ell}}{\tau^2_{r\ell}}$$

where $\Delta_{r\ell}$ is the penalty matrix of nonparametric effect $(r\ell)$ based on a random walk prior and mean

$$\mu_{\beta_{r\ell}} = P^{-1}_{\beta_{r\ell}}X^t_{r\ell}\Sigma^{-1}_{r|-r}(y_r - \widetilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r}))$$

where $\widetilde{\eta}_r = \eta_r - f_{r\ell}$ when $f_{r\ell}$ is to be estimated. Further, $\mathrm{E}(\varepsilon_r|\varepsilon_{-r})$ with $\varepsilon_r = (\varepsilon_{r11}, \ldots, \varepsilon_{rnn_n})^t$ is the conditional mean of the error terms with

$$E(\varepsilon_{rij}|\varepsilon_{-r,ij}) = \mu_{rij} + \frac{\sigma_{12,ij}}{\sigma^2_{-r,ij}}(y_{-r,ij} - \mu_{-r,ij} - \eta_{-r,ij})$$

and $\Sigma_{r|-r}$ is the conditional covariance matrix with $\Sigma_{r|-r} = \mathrm{diag}(\sigma^2_{(r|-r),11}, \ldots, \sigma^2_{(r|-r),nn_n})$ and

$$\sigma^2_{(r|-r),ij} = \mathrm{Var}(\varepsilon_{rij}|\varepsilon_{-r,ij}) = \sigma^2_{rij} - \frac{\sigma^2_{12,ij}}{\sigma^2_{-r,ij}}.$$

Note that the posterior mean of some function $f_{r\ell}$ is given by (subject to centering constraints)

$$f_{r\ell}(\cdot) = (X^t_{r\ell}\Sigma^{-1}_{r|-r}X_{r\ell} + \frac{1}{\tau^2_{r\ell}}\Delta_{r\ell})^{-1}X^t_{r\ell}\Sigma^{-1}_{r|-r}(y_r - \widetilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r})).$$

Here, it can be easily seen that the DPM prior induces different variances and therefore $\Sigma_{r|-r}$ weighs observations accordingly just as in the case of heteroscedasticity.
The full conditionals for the smoothing variance parameters $\tau^2_{r\ell},\ \ell = 1, \ldots, p_r,\ r = 1, 2$ follow inverse Gamma distributions

$$\tau^2_{rl}|\cdot \sim IG(a'_{\tau_{r\ell}}, b'_{\tau_{r\ell}})$$

with parameters

$$a'_{\tau_{r\ell}} = a_{\tau_{r\ell}} + \frac{\text{rank}(\Delta_{r\ell})}{2}, \quad b'_{\tau_{r\ell}} = b_{\tau_{r\ell}} + \frac{1}{2}\beta^t_{r\ell}\Delta_{r\ell}\beta_{r\ell}.$$

Parametric effects The full conditionals for the coefficients γ_r of parametric effects are Gaussian

$$\gamma_r|\cdot \sim \text{N}(\mu_{\gamma_r}, P^{-1}_{\gamma_r})$$

$$\begin{aligned} \text{with precision matrix} \qquad & P_{\gamma_r} = V_r^t \Sigma^{-1}_{r|-r} V_r \\ \text{and mean} \qquad & \mu_{\gamma_r} = P^{-1}_{\gamma_r} V_r^t \Sigma^{-1}_{r|-r}(y_r - \widetilde{\eta}_r - \text{E}(\varepsilon_r|\varepsilon_{-r})) \end{aligned}$$

where $\widetilde{\eta}_r = \eta_r - V_r\gamma_r$.

Components of the error distribution In our default implementation, we make use of R function DPdensity (Jara, Hanson, Quintana, Müller & Rosner, 2011) for error density estimation adopting algorithm 8 of Neal (2000) with one temporarily existing auxiliary parameter. In the following, the full conditionals are summarized, for more details see Neal (2000).

- Let $c_i \in \{1, \ldots, K^*\}, i = 1, \ldots, n$ indicate the cluster observation i belongs to.

  For $i = 1, \ldots, n$:

  - If $c_i = c_h$ for some $h \neq i$, create auxiliary component c^* with $(\mu_{c^*}, \Sigma_{c^*})$ drawn from G_0.
  - If $c_i \neq c_h$ for all $h \neq i$, let $c^* = c_i$ with $(\mu_{c^*}, \Sigma_{c^*}) = (\mu_{c_i}, \Sigma_{c_i})$.

– Draw a new value for c_i using

$$c_i|c_{-i}, y_{1i}, y_{2i}, \mu_1, \Sigma_1, \ldots, \mu_{K^*}, \Sigma_{K^*}, \mu_{c^*}, \Sigma_{c^*} \sim b\sum_{l=1}^{k^-} \frac{n_l^{-i}}{n-1+\alpha} F\left((\varepsilon_{1i}, \varepsilon_{2i}), \mu_l, \Sigma_l\right) + b\frac{\alpha}{n-1+\alpha} F\left((\varepsilon_{1i}, \varepsilon_{2i}), \mu_{c^*}, \Sigma_{c^*}\right)$$

where k^- is the number of distinct c_h for $h \neq i$, n_l^{-i} is the number of c_h for $h \neq i$ that are equal to l, b is a normalizing constant and $F\left((\varepsilon_{1i}, \varepsilon_{2i}), \mu_l, \Sigma_l\right)$ the likelihood for observation i.

- Discard those μ_l, Σ_l that are not associated with one or more observations.
- For all $l \in \{c_1, \ldots, c_n\}$: Update μ_l and Σ_l using $\mu_l|\cdot \sim \mathrm{N}(m_{\mu_l}, P_{\mu_l}^{-1})$ and $\Sigma_l|\cdot \sim \mathrm{IW}(s'_\Sigma, S'_\Sigma)$ with

$$\begin{aligned}
m_{\mu_l} &= (\tau_\Sigma + 1)^{-1}\left(\tau_\Sigma \mu_0 + \sum_{i:c_i=l} ((y_{1i}, y_{2i}) - (\eta_{1i}, \eta_{2i}))^t\right) \\
P_{\mu_l}^{-1} &= \frac{\tau_\Sigma^{-1}}{1+\tau_\Sigma^{-1}}\Sigma_l/n_l = (\tau_\Sigma+1)^{-1}\Sigma_l/n_l, \\
s'_\Sigma &= s_\Sigma + \frac{n_l}{2} \\
S'_\Sigma &= S_\Sigma + \frac{1}{2}\frac{1}{1+\tau_\Sigma^{-1}} \sum_{i:c_i=l} ((y_{1i}, y_{2i}) - (\eta_{1i}, \eta_{2i}) - \mu_0)^t ((y_{1i}, y_{2i}) - (\eta_{1i}, \eta_{2i}) - \mu_0)
\end{aligned}$$

- In case τ_Σ is not fixed, the full conditionals of τ_Σ are

$$\tau_\Sigma \sim \mathrm{Ga}\left(\frac{a_\Sigma + K^*}{2},\ \frac{1}{2}\left(b_\Sigma + \sum_{l=1}^{K^*} \Sigma_l^{-1}(\mu_l - \mu_0)^2\right)\right)$$

- The concentration parameter α in case of a gamma prior is drawn from a mixture of two gamma distributions

$$\alpha|\cdot \sim \frac{a_\alpha + K^* - 1}{n(b_\alpha - \log\omega)} \operatorname{Ga}(a_\alpha + K^*, b_\alpha - \log\omega) + \left(1 - \frac{a_\alpha + K^* - 1}{n(b_\alpha - \log\omega)}\right) \operatorname{Ga}(a_\alpha + K^* - 1, b_\alpha - \log\omega)$$

where ω is a latent variable sampled from a beta distribution $\omega \sim \text{Be}(\alpha + 1, n)$.

In case of a discrete prior for α as in Conley, Hansen, McCulloch & Rossi (2008), α is drawn from a multinomial distribution. See Conley, Hansen, McCulloch & Rossi (2008) for details.

4.3.3 Smoothing Parameter Estimation

In general, all nonparametric smoothing techniques involve a smoothing parameter controlling the roughness of the fit, may it be the bandwidth in kernel smoothing, the number of knots or components in regression splines or series estimators or a parameter controlling the impact of some penalization term in penalized splines smoothing. This smoothing parameter has a strong impact on the estimate and has to be carefully chosen in the finite sample context. However, data-driven choice is rather overlooked in many theoretical works on nonparametric instrumental variable estimators focusing on asymptotic properties.

In the control function approach, smoothing parameter choice for the control function $\text{E}(\varepsilon_2|\varepsilon_1)$ and of the covariate functions have to be addressed differently. Here, smoothing parameter choice is even more problematic, since smoothness of functions in the first stage and of the control function influence the way of endogeneity bias correction for $f_{21}(y_1)$. Thereby, the major problem is to find the smoothing parameter for the control function. Given this smoothing parameter is correctly chosen, it seems plausible that the remaining ones can be found using common criteria like cross-validation. Newey, Powell & Vella (1999) minimize the cross-validation (CV) criterion over a multidimensional grid and thus treat the control function in the same way as $f_{21}(y_1)$. That is, the MSE of the additive predictor as a whole is (asymptotically) minimized instead of the MSE of $f_{21}(y_1)$ given $\text{E}(\varepsilon_2|\varepsilon_1)$. Marra & Radice (2011) take the same route using penalized splines with quadratic roughness penalties and minimize a multivariate version of generalized cross-validation (GCV). In Section 4.4.2, we show that this can lead to a confounded estimate of $f_{21}(y_1)$ due to inappropriate choices for the smoothing parameter of the control function. Choosing the smoothing parameter from a global optimization criterion often induces insufficient smoothness, although situations with oversmoothing may also

occur. In general, global optimization criteria are not suitable for determining smoothing parameters that minimize the MSE of $f_{21}(y_1)$.
Su & Ullah (2008) propose a "plug-in" estimator for the smoothing parameter in a multidimensional function $f(y_1, \varepsilon_1)$ (in the model with $q_1 = q_2 = p_2 = 0,\ p_1 = 1$) where $f(\cdot, \cdot)$ is a two-dimensional function using kernel regression with a product kernel with single bandwidth, and a pilot bandwidth for estimating $\hat{f}_1(z_1)$. Here, choosing the pilot bandwidth and the assumption of a single bandwidth for $f(y_1, \varepsilon_1)$ might be problematic.
Our Bayesian approach is closely related to the control function approach. For comparison with Equation (4.2), consider the conditional distribution of y_2 given y_1, then

$$y_{2i} = \gamma_{20} + f_{21}(y_{1i}) + \sum_{\ell=1}^{q_2} x_{2\ell i}\gamma_{2\ell} + \sum_{\ell=1}^{p_2} f_{2,\ell+1}(z_{2\ell i}) + \mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i}) + \xi_i, \quad \xi_i \sim \mathrm{N}(0, \sigma^2_{(2|1),i})$$

with conditional variance $\sigma^2_{(2|1),i} = \sigma^2_{2,i} - \frac{\sigma^2_{12,i}}{\sigma^2_{1,i}}$ and "control function" $v(\varepsilon_{1i}) = \mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i}) = \mu_{2i} + \frac{\sigma_{12,i}}{\sigma^2_{1,i}}(\varepsilon_{1i} - \mu_{1i})$. Estimates for parameters in $\mathrm{E}(\varepsilon_{2i}|\varepsilon_{1i})$ result from the DP mixture and covariate effects $f_{2\ell}(\cdot)$ are estimated by penalized splines. Compared to parametric frequentist approaches and Bayesian approaches assuming bivariate normality, $\frac{\sigma_{12,i}}{\sigma^2_{1,i}}$ may vary with observation i rather than being constant. This formulation of the conditional mean of the error terms also shows that in the presence of heteroscedasticity, endogeneity bias correction may fail when bivariate normality with constant variance is assumed.
Compared to nonparametric frequentist approaches, $\frac{\sigma_{12,i}}{\sigma^2_{1,i}}$ acts like a varying coefficient allowing the degree of endogeneity correction to be different over observations. The nonconstant variances $\sigma^2_{1,i}$ and means μ_{1i} shrink the error terms of the first stage equation towards their (nonconstant) mean and thereby automatically down weight outliers in ε_{1i}. Here, on the one hand the "smoothing parameter" is the number of mixture components governed by the data and prior on the concentration parameter α. On the other hand, τ_Σ plays an important role for the smoothness of the error density. As mentioned before, a small τ_Σ allows the μ_{1i} to vary more strongly around its mean which translates in a possibly stronger downweighting of outliers in ε_{1i} depending on τ_Σ. Note that control function approaches can be extremely sensitive to outliers in the error distribution if these are not explicitly handled, since they do not account for the high variability of the control function at extreme values of ε_1 (outliers) where observations are scarce. Performance of the DPM approach in case of residual outliers and capability of explaining unobserved heterogeneity will be investigated in Section 4.4.2. However, note that there is no such thing as a free lunch and the downweighting of outliers can also turn into a disadvantage

in specific situations. If y_1 or y_2 are discrete and concentrated very strongly on only a few numbers, rarer measurements may be misinterpreted as outliers and variability can then completely be explained by the error distribution leaving no variation to be explained by the covariates (in particular in case of binary covariates). We observed this problem in a re-analysis of the relationship between years of education (as discrete endogenous covariate) and wages in the US (Card, 1995) with nonparametric effect of the control variable age (or transformations thereof). Here, half of the observed number of years of schooling were 12 and 16 (corresponding to usual years of schooling in the US education system) resulting in an extremely imbalanced weighting of the observations. In the present example, the omitted variable "education system" can be understood as inducing unobserved heterogeneity (clustering at 12 and 16 years of schooling is unexplained by the included covariates) which is then absorbed by the predicted error terms leaving little variation to be explained by the remaining explanatory variables. Note that this issue is not specific to our proposed approach but applies to all regression approaches with DPM prior on the error density as in Chib & Greenberg (2010) and in Leslie, Kohn & Nott (2007). A rough diagnostic check is to check the estimated error density for discreteness. In this case, estimates should be treated with caution.
Note that in contrast to the frequentist approaches, we do not impose dependencies between values of $v(\varepsilon_{1i})$ for adjacent ε_{1i} and $\frac{\sigma_{12,i}}{\sigma^2_{1,i}}$ is also not a function of ε_1. Also note that the DP prior specification allows "for different degrees of smoothing across the sample space through the use of possibly differing variances" (Escobar & West, 1995) and thus the "smoothing parameter" of the conditional mean can be considered to be locally adaptive. See Escobar & West (1995) for connections between DPM and kernel density estimation with varying bandwidth.
The smoothness of functions $f_{r\ell}(\cdot)$ is controlled by the smoothing variance $\tau^2_{r\ell}$ which acts like an inverse smoothing parameter to which a prior distribution is assigned and which is thus also prior-data driven. Since weakly informative priors for $\tau^2_{r\ell}$ are chosen, the degree of smoothness chosen is generally quite insensitive against hyperparameter choices as shown in Lang & Brezger (2004) for the single equation case.

4.3.4 Simultaneous Bayesian Credible Bands

The Bayesian counterpart to simultaneous confidence bands are simultaneous credible bands. Simultaneous inference is important in order to assess the estimation uncertainty for the entire curve allowing us to make statements about the significance of an effect or feature significance and to perform specification tests. While a frequentist $(1-\alpha)100\%$ simultaneous confidence band is defined such that in case of multiple replications of the

data with the same mean function, $(1-\alpha)100\%$ of the estimated functions will be *entirely* inside the band, a Bayesian simultaneous credible band is defined as the region I_α such that $P_{f|Y}(f \in I_\alpha) = 1 - \alpha$, i.e. the posterior probability that the *entire* true function f is inside the region given the data equals to $1 - \alpha$. Note that the commonly used (frequentist) pointwise bands usually only provide that *on average* $(1 - \alpha)100\%$ of the data points of the true function are inside the band (in an experiment where the data is sampled with the same f many times).
In general, Bayesian simultaneous credible bands are slightly broader than the frequentist simultaneous confidence bands (Krivobokova, Kneib & Claeskens, 2010). This is partially explained by the different construction where the level can not be interpreted in the usual frequentist way but can also partially be attributed to the fact that uncertainty about all model parameters is appropriately reflected in the Bayesian credible intervals. However, in the instrumental variable regression context, their advantage is that they naturally incorporate uncertainty from the estimation of all the unknowns in the model including those of the "first stage" equation explaining the endogenous covariate, which is particularly difficult in the frequentist framework. Even uncertainty due to estimating the corresponding smoothing parameters is taken into account. Moreover, no hard-to-find asymptotic distribution of the estimator is necessary as in the frequentist framework and we do not have to make any distributional assumption, i.e. also asymmetric bands can be obtained.
We follow Krivobokova, Kneib & Claeskens (2010) and obtain Bayesian simultaneous credible bands from scaling the pointwise credible intervals derived from the $\alpha/2$ and $1 - \alpha/2$ quantiles of the function samples from the MCMC output with a constant factor until $(1 - \alpha)100\%$ of the sampled curves are contained in the credible band. Thereby, the information on the possibly nonnormal error distribution is preserved and the complete variability is taken into account without overly demanding computationally effort. Simulations in Section 4.4.2 show that they preform very well even in rather small samples and complex settings.

4.4 Simulations

4.4.1 Parametric Model

Settings

In this section, settings with linear covariate effects are simulated with the following goals: First, the Bayesian approach is compared to the well-established two-stage least squares estimator showing that it is capable of correcting endogeneity bias. Second, it is

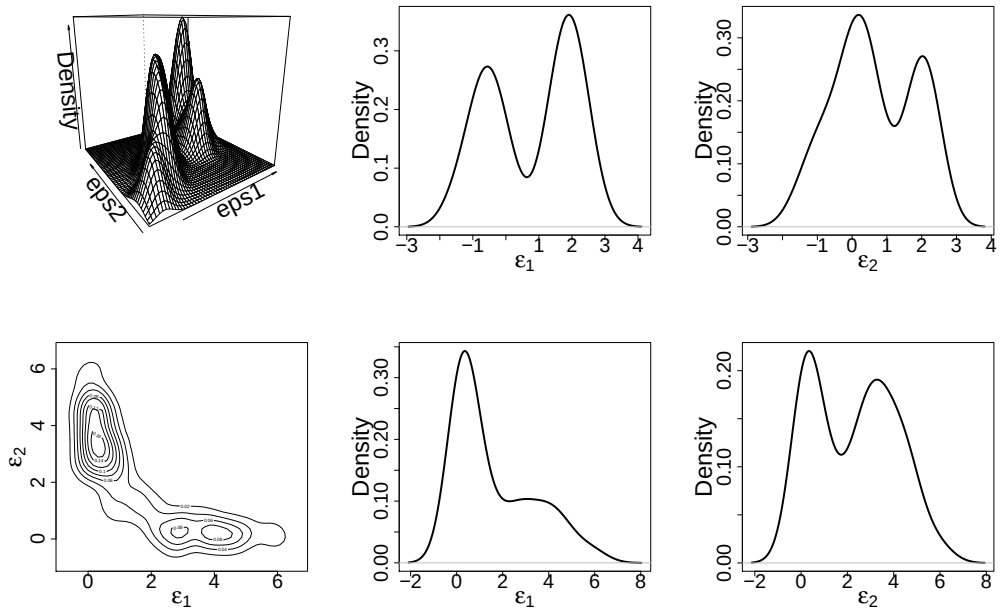

Figure 4.1: Joint and marginal densities in one Monte Carlo draw of simulation setting (iii) (top panels) and setting (iv) (bottom panels).

shown that in certain unfavorable situations, more precisely in the cases of outliers in the error distribution and nonlinear conditional means, the Bayesian approach outperforms the 2SLS procedure. Thus, this section supplements the studies of Conley, Hansen, McCulloch & Rossi (2008) where normal and log-normal error distributions were simulated. Note however, that in their settings comparisons to 2SLS have to be treated with caution since they use settings with 10 instruments (the "many instruments" case), where 2SLS is known to be inconsistent (as given in their own appendix).
In all parametric settings, we consider the following basic model

$$\begin{aligned} y_2 &= y_1 + z_2 + \varepsilon_2 \\ y_1 &= z_1 + z_2 + \varepsilon_1 \end{aligned}$$

where z_2 and z_1 are independently uniformly distributed on $[0, 1]$ and all coefficients are equal to 1. We consider four different bivariate distributions for the error terms:

(i) a simple bivariate normal distribution with a quite high degree of endogeneity

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \right).$$

(ii) a mixture of two normal distributions that adds outliers (with very small correlation $\rho = 0.1$) on (i):

$$\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \sim 0.95\,\mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix} \right) + 0.05\,\mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 5 & 0.5 \\ 0.5 & 5 \end{pmatrix} \right).$$

(iii) a mixture of four bivariate normals with weights 0.3, 0.2, 0.3 and 0.2, means $(2,2)^t$, $(1.5,0.5)^t$, $(-0.3,0)^t$ and $(-1,-1)^t$, all variances (in each mixture components and both equations) equal to 0.1 and correlations 0.5, 0.2, 0.6 and 0.8 between the equations. This setting is an example of (unobserved) heterogeneity with varying degrees of endogeneity in each cluster. Densities of an example draw are shown in Figure 4.1.

(iv) a symmetric bivariate distribution which is conditionally normal with nonlinear conditional mean, i.e. $\varepsilon_1|\varepsilon_2 \sim \mathrm{N}\left(\frac{4}{\varepsilon_2^2+1}, \frac{1}{\varepsilon_2^2+1}\right)$ and vice versa for $\varepsilon_2|\varepsilon_1$ (Meng & Gelman, 1991). Note that the degree of endogeneity varies over observations. Densities of an example draw are shown in Figure 4.1.

Obviously, the strength of the instruments as well as the degrees of endogeneity vary over the settings. In each setting, we simulated 500 Monte Carlo replications with rather small and moderately large sample sizes $n = 100,\ 400$. For our DPM approach, the initial 3000 iterations are discarded for burn-in and every 30th iteration of the subsequent 30.000 iterations is used for inference. As discussed in Section 4.2.4, we choose a weakly informative prior on the error distribution with $s_\Sigma = 3$, $S_\Sigma = \mathrm{diag}(0.2, 0.2)$, $\mu_0 = (0,0)^t$ and $a_\alpha = b_\alpha = 2$. Labeled as "DPM1", we consider first a fixed τ_Σ chosen according to Conley, Hansen, McCulloch & Rossi (2008)'s assessment strategy based on the observation that the errors are marginally t-distributed and thus $\mu_r \sim \sqrt{S_{\Sigma_{rr}}/\tau_\Sigma(s_\Sigma - 1)}\,\mathrm{t}_{s_\Sigma - 1}$. Considering that the data were centered and standardized, τ_Σ is then chosen such that $P(-10 < \mu_r < 10) = 0.8$ which results in $\tau_\Sigma = 0.036$ given $s_\Sigma = 3$ and $S_\Sigma = 0.2 I_2$. Second, we consider a weakly informative gamma distribution for τ_Σ with $a_\Sigma = 1$ and $b_\Sigma = 100$, labeled as "DPM2" in the following.

Table 4.1: Parametric simulation setting (i): Bivariate normality.

	point estimates				confidence intervals			
	mean bias	median bias	RMSE	IQR	coverage	ave.width	med.width	rej. rate
$n = 100$								
OLS	0.65	0.64	0.65	0.09	0.00	0.29	0.28	1.00
2SLS	-0.18	-0.02	0.79	0.53	0.93	3.22	1.47	0.61
DPM1	-0.08	-0.05	0.40	0.49	0.97	2.03	1.73	0.43
DPM2	-0.03	-0.01	0.32	0.42	0.97	1.76	1.55	0.51
$n = 400$								
OLS	0.65	0.65	0.65	0.05	0.00	0.14	0.14	1.00
2SLS	-0.01	0.00	0.18	0.24	0.96	0.72	0.69	0.98
DPM1	-0.04	-0.03	0.19	0.25	0.97	0.80	0.75	0.91
DPM2	-0.04	-0.02	0.19	0.24	0.96	0.78	0.74	0.94

Results

In simulation settings (i) (Table 4.1) and $n = 100$, the DPM approach performs overall better than 2SLS especially in terms of variability of the point estimates. Particularly, the RMSEs (evaluated at the design points) are considerably lower for the DPM approach. Note that 6.6% of the 2SLS estimates even had a negative sign (versus virtually none in the DPM approach with 0.6% and 0.2%, respectively). In setting (i), the DPM approach with gamma prior on τ_Σ performs only slightly better than with fixed τ_Σ. This becomes more pronounced in setting (ii) (Table 4.2, $n = 100$), however, where in presence of outliers, assigning a hyperprior is clearly preferable. While RMSEs of 2SLS increase in presence of outliers in setting (ii), this was not the case for the DPM estimator. For the larger sample size $n = 400$, 2SLS and the DPM approach perform almost identically well and as expected, the impact of the prior on τ_Σ diminishes. Note that in settings (i) and (ii) instruments are very weak with a population R^2 of $R^2_{pop} = \frac{\text{Var}(z_1)}{\text{Var}(z_1)+\text{Var}(z_2)+\sigma_1^2} = \frac{1/12}{1/12+1/12+\sigma_1^2} \approx 0.07$ and even slightly lower in setting (ii).

In settings (iii) and (iv) (Tables 4.3 and 4.4), both actually examples of nonlinear conditional residual means, bias, RMSE and IQR of the 2SLS estimators are excessively large in the case of $n = 100$ while those of the DPM estimator are considerably lower. Due to the strongly increased widths of the 2SLS confidence intervals, coverage probability of the intervals are, however, still close to the nominal level. Still, this also has an impact on the power of detecting a significant positive effect: On a 5% level, rejection rates of 59% and 32% for the 2SLS estimator for settings (iii) and (iv), respectively, were observed versus 100% for the DPM estimator. In these two settings, the DPM estimator with fixed τ_Σ performed best, since estimation of τ_Σ increased variability in DPM2. Here,

Table 4.2: Parametric simulation setting (ii): Bivariate normality with outliers.

	point estimates				confidence intervals			
	mean bias	median bias	RMSE	IQR	coverage	ave.width	med.width	rej. rate
$n = 100$								
OLS	0.55	0.56	0.56	0.16	0.00	0.32	0.32	1.00
2SLS	-0.00	-0.01	3.10	0.56	0.94	93.40	1.55	0.59
DPM1	-0.06	-0.04	0.39	0.46	0.96	2.04	1.79	0.42
DPM2	0.01	0.03	0.31	0.42	0.95	1.57	1.37	0.63
$n = 400$								
OLS	0.54	0.55	0.55	0.08	0.00	0.16	0.16	1.00
2SLS	-0.01	0.01	0.20	0.26	0.95	0.80	0.76	0.96
DPM1	-0.04	-0.01	0.19	0.24	0.96	0.80	0.76	0.93
DPM2	-0.03	-0.01	0.18	0.23	0.96	0.77	0.74	0.95

Table 4.3: Parametric simulation setting (iii): Mixture of bivariate normals (unobserved clusters).

	point estimates				confidence intervals			
	mean bias	median bias	RMSE	IQR	coverage	ave.width	med.width	rej. rate
$n = 100$								
OLS	0.77	0.77	0.77	0.06	0.00	0.17	0.17	1.00
2SLS	-0.50	0.01	10.41	0.50	0.92	193.46	1.55	0.59
DPM1	0.11	0.11	0.16	0.17	0.92	0.61	0.60	1.00
DPM2	0.12	0.13	0.18	0.18	0.93	0.69	0.68	1.00
$n = 400$								
OLS	0.77	0.77	0.77	0.02	0.00	0.08	0.08	1.00
2SLS	-0.04	-0.00	0.24	0.26	0.94	0.90	0.79	0.89
DPM1	0.03	0.03	0.07	0.08	0.94	0.27	0.26	1.00
DPM2	0.03	0.04	0.07	0.09	0.95	0.28	0.28	1.00

also for $n = 400$, due to the nonlinear conditional means, the DPM approach performs better than 2SLS in terms of efficiency (MSE and IQR) and interval widths. Again, the importance of the prior on τ_Σ diminishes for increasing sample size.

4.4.2 Nonparametric Model

Settings

In our first two settings with nonparametric covariate effects, we replicate DGPs 1 and 4 of Su & Ullah (2008) aiming at getting some insight into the comparison of our Bayesian approach with Pinkse (2000)'s, Newey & Powell (2003)'s and Su & Ullah (2008)'s ap-

Table 4.4: Parametric simulation setting (iv): Nonlinear conditional mean.

	point estimates				confidence intervals			
	mean bias	median bias	RMSE	IQR	coverage	ave.width	med.width	rej. rate
$n = 100$								
OLS	-0.82	-0.82	0.82	0.07	0.00	0.22	0.22	0.87
2SLS	-0.81	-0.08	16.66	0.79	0.91	557.31	2.24	0.32
DPM1	-0.05	-0.05	0.14	0.18	0.94	0.57	0.56	1.00
DPM2	-0.06	-0.07	0.15	0.20	0.98	0.74	0.72	1.00
$n = 400$								
OLS	-0.81	-0.81	0.81	0.04	0.00	0.11	0.11	1.00
2SLS	0.06	-0.04	0.38	0.38	0.93	1.45	1.09	0.94
DPM1	-0.02	-0.02	0.07	0.10	0.95	0.27	0.27	1.00
DPM2	-0.03	-0.03	0.08	0.10	0.96	0.31	0.31	1.00

proaches. Moreover, we compare our results with Marra & Radice (2011)'s approach (extending the control function approach of Newey, Powell & Vella (1999) to penalized splines). More precisely, we consider settings

(a) DGP1 of Su & Ullah (2008):

$$\begin{aligned} y_2 &= \log(|y_1 - 1| + 1)\text{sgn}(y_1 - 1) + \varepsilon_2 \\ y_1 &= z_1 + \varepsilon_1 \end{aligned}$$

with $z_1 \overset{\text{i.i.d.}}{\sim} \text{N}(0,1)$ and $\begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \end{pmatrix} \overset{i.i.d}{\sim} \text{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \theta \\ \theta & 1 \end{pmatrix}\right)$.

(b) DGP4 of Su & Ullah (2008):

$$\begin{aligned} y_2 &= 2\Phi(y_1) + \varepsilon_2 \\ y_1 &= \log(0.1 + z_1^2) + \varepsilon_1 \end{aligned}$$

with $\Phi(\cdot)$ the cdf of the standard normal and $\varepsilon_2 = \theta w + 0.3v_2$, $\varepsilon_1 = 0.5w + 0.2v_1$ and $z_{1i} = 1 + 0.5z_{1,i-1} + 0.5v_z$. w, v_1, v_2 and v_z are i.i.d. sums of 48 independent random variables each uniformly distributed on $[-0.25, 0.25]$ and thus according to the central limit theorem nearly standard normal but with compact support $[-12, 12]$.

In settings (b.ii) and b.iii) the error distribution in (b) is replaced by the distributions in settings (ii) and (iii) of the previous section, respectively. In setting (b.v), the distribution

in (b) is replaced by one of the distributions given in Marra & Radice (2011) which exactly resembles the structural assumptions of the control function approach:

$$\varepsilon_1 = g_1(w) + v_1 \qquad \varepsilon_2 = g_2(w) + v_2$$

with $w \sim U(0,1)$, $g_1(w) = -\exp(-3w)$ and $g_2(w) = -0.5(w + \sin(\pi x^2.5))$ standardized to have variance one and $v_1, v_2 \overset{\text{i.i.d.}}{\sim} \text{N}(0,1)$. Note that in settings (b) and (b.v), w can be considered as an omitted variable with linear and nonlinear effects, respectively.
Again, 500 Monte Carlo replications with $n = 100,\ 400$ are considered. For the Bayesian approach, we use a burn-in of 5.000 iterations and use 1.000 of the subsequent 40.000 iterations for estimation. Further, cubic B-splines based on 25 and 40 knots for sample sizes of 100 and 400, respectively, and a second-order random walk prior were used for the Bayesian P-splines.

Results

In Table 4.5, mean RMSEs and coverage rates of 95% simultaneous credible bands (when available) for DGP 1 and 4 of Su & Ullah (2008) (settings (a) and (b)) are given. We compare naive (i.e. without bias correction) estimation using local linear regression (with normal kernel) and LSCV smoothing parameter selection (as Su & Ullah (2008) did) and the two step control function approach using penalized splines (we used cubic B-splines with second order difference penalty and same number of knots as for the DPM approach) with GCV smoothing parameter selection following Marra & Radice (2011) to our DPM approach (with hyperparameter settings DPM1 and DPM2 as in the previous subsection). As a benchmark, we give the results for the models using the true but unobserved $y_2 - \text{E}(\varepsilon_2|\varepsilon_1)$ as response.
We find RMSEs for all estimators that are considerably smaller than those given in Su & Ullah (2008). Note that we even obtained better results for the naive estimator using LSCV. This is most probably due to the fact that while we used a numerical minimization algorithm with a random starting value to minimize the LSCV criterion, Su & Ullah (2008) (personal communication) chose the bandwidth h according to $h = c\sqrt{\text{Var}(y_1)}n^{-1/5}$ with a limited grid search over c. Thereby, they obtained RMSEs that only slightly changed with increasing degree of endogeneity which is rather implausible. While both the control function and DPM approach decreased the mean RMSE compared to the naive estimator, the DPM approach performed slightly better with negligible impact of the prior choice.
Table 4.6 gives results for settings (b.ii), (b.iii) and (b.v). In settings (b.ii) and (b.iii) (outliers and multimodal error density, unobserved heterogeneity) the control function

Table 4.5: Setting (a) and (b): DGPs of Su & Ullah (2008)

		DGP1				DGP4			
		n=100		n=400		n=100		n=400	
θ		RMSE	coverage	RMSE	coverage	RMSE	coverage	RMSE	coverage
0.2	naive (LSCV)	0.242	–	0.183	–	0.154	–	0.136	–
	naive (Bayes)	0.228	0.912	0.173	0.766	0.145	0.466	0.131	0.012
	DPM1	0.213	0.980	0.117	0.978	0.075	0.976	0.042	0.988
	DPM2	0.213	0.982	0.117	0.980	0.075	0.972	0.042	0.980
	CF with GCV	0.242	–	0.129	–	0.087	–	0.045	–
	benchmark (GCV)	0.211	–	0.117	–	0.067	–	0.038	–
	benchmark (Bayes)	0.182	0.980	0.105	0.988	0.061	0.992	0.037	0.992
0.5	naive (LSCV)	0.395	–	0.361	–	0.336	–	0.322	–
	naive (Bayes)	0.389	0.408	0.365	0.000	0.331	0.010	0.318	0.000
	DPM1	0.206	0.970	0.113	0.982	0.108	0.968	0.058	0.982
	DPM2	0.207	0.968	0.113	0.978	0.108	0.968	0.058	0.976
	CF with GCV	0.231	–	0.122	–	0.127	–	0.064	–
	benchmark (GCV)	0.188	–	0.105	–	0.067	–	0.038	–
	benchmark (Bayes)	0.165	0.968	0.094	0.988	0.061	0.992	0.037	0.992
0.8	naive (LSCV)	0.585	–	0.564	–	0.519	–	0.505	–
	naive (Bayes)	0.582	0.002	0.571	0.000	0.521	0.002	0.507	0.000
	DPM1	0.186	0.960	0.100	0.974	0.149	0.960	0.079	0.974
	DPM2	0.187	0.958	0.100	0.970	0.149	0.962	0.079	0.970
	CF with GCV	0.209	–	0.106	–	0.175	–	0.090	–
	benchmark (GCV)	0.138	–	0.076	–	0.067	–	0.038	–
	benchmark (Bayes)	0.122	0.976	0.069	0.984	0.061	0.992	0.037	0.992

approach is clearly outperformed by the DPM approach. Figure 4.2 shows the estimated curves in the first 50 simulation runs of setting (b.iii) illustrating that estimates of the control function approach can be seriously confounded when $\mathrm{E}(\varepsilon_2|\varepsilon_1)$ is not a smooth function. Clearly, this cannot be only attributed to the higher variability of the cross-validated smoothing parameter of $\hat{f}_{21}(y_1)$. Also in setting (b.v), the DPM approach performs better although not as pronounced.

In all settings, the DPM approach provides simultaneous credible bands with frequentist coverage rates above the nominal level. That is, the credible bands were successful in taking into account all the variability in the estimation. On the other hand, the credible bands are slightly conservative in a frequentist coverage sense which is unsurprising since this is a well-known property of Bayesian credible bands also observed in Krivobokova, Kneib & Claeskens (2010) in the single equation case. Note that for the control function approach as well as for the approaches compared in Su & Ullah (2008), no simultaneous confidence bands are available.

In summary, the proposed approach outperformed the control function approach based

Table 4.6: Settings (b.ii), (b.iii) and (b.v): More complex distributions

		(b.ii): Outliers		(b.iii): Mixture Distribution		(b.v): Omitted Variable	
n		RMSE	coverage	RMSE	coverage	RMSE	coverage
100	naive (Bayes)	0.610	0.030	0.922	0.000	0.634	0.084
	DPM1	0.268	0.976	0.124	0.980	0.395	0.958
	DPM2	0.262	0.974	0.121	0.974	0.393	0.962
	CF with GCV	0.409	–	0.339	–	0.435	–
	benchmark (GCV)	0.258	–	0.063	–	0.226	–
	benchmark (Bayes)	0.213	0.836	0.060	0.990	0.195	0.974
400	naive (Bayes)	0.580	0.000	0.926	0.000	0.616	0.000
	DPM1	0.154	0.974	0.059	0.982	0.228	0.940
	DPM2	0.153	0.974	0.058	0.982	0.224	0.938
	CF with GCV	0.355	–	0.163	–	0.243	–
	benchmark (GCV)	0.196	–	0.034	–	0.128	–
	benchmark (Bayes)	0.142	0.742	0.034	0.994	0.115	0.974

on GCV smoothing parameter selection and the estimators of Pinkse (2000), Newey & Powell (2003) and Su & Ullah (2008) (relying on the results given in Su & Ullah, 2008). This shows the extreme importance of the smoothing or tuning parameter which can hardly be estimated in the frequentist approaches. Moreover, only our Bayesian approach provided us with simultaneous credible bands which performed extremely well even in the case of rather complex error distributions and small sample sizes.

4.5 Application: Class Size Effects on Student Achievements

In a very influential paper, Angrist & Lavy (1999) analyzed the effect of class size on 4th and 5th grades students tests scores in Israel. Their main analysis relies on 2SLS using a specific instrumental variable in the context of a linear regression model with random effects. More precisely, among others they consider the model

$$tscore_{ji} = \gamma_{20} + \gamma_{21} csize_{ji} + \gamma_{22} disadv_{ji} + \nu_j + \varepsilon_{2ji}$$

where $tscore_{ji}$ is the class level average of a reading comprehension test score, $csize_{ji}$ the number of students and $disadv_{ji}$ the fraction of disadvantaged students in class i of school j, respectively. Further, ν_j is a school-specific random effect.
As discussed in Angrist & Lavy (1999), endogeneity of $csize_{ji}$ due to non-random assignment of class sizes complicates estimation of the class size effect. To deal with the endogeneity of $csize_{ji}$, Angrist & Lavy (1999) exploit an exogenous assignment rule based

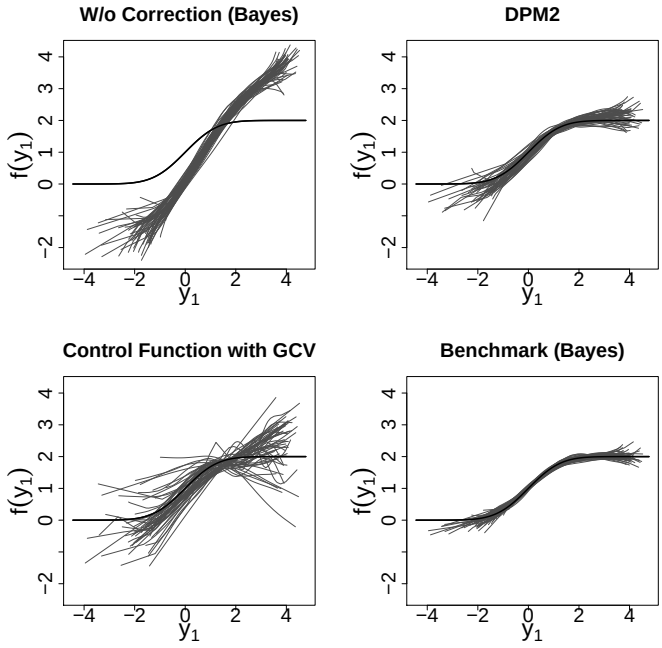

Figure 4.2: Setting (b.iii): Estimated curves in first 50 simulation runs for $n = 100$.

on governmental recommendations of 40 students as the maximum class size. That is, they define the predicted class size $pcsize_{ji}$ of class j in school i as an instrument given by

$$pcsize_{ji} = \frac{enrol_j}{\text{int}[(enrol_j - 1)/40] + 1},$$

where $enrol_j$ is the beginning of the year enrollment in school j for a given grade and $\text{int}(k)$ is the largest integer less or equal to k. *pcsize* implies the rule that schools facing an enrollment size less or equal to 40 must have only one class. Similarly, schools with enrollment between 41 and 80 must accommodate students in two classes, and so on. Using a sample of 2019 public schools and assuming a first stage equation

$$csize_{ji} = \gamma_{10} + \gamma_{11} pcsize_{ji} + \gamma_{12} disadv_{ji} + \varepsilon_{1ji}$$

they fit the model using 2SLS and find, for fourth and fifth graders, class size effects of -0.110 and -0.158, respectively, with standard errors of 0.040 each resulting in the conclusion of a significantly negative effect on the reading comprehension test score. When

applying our DPM approach to the parametric model specification, i.e. when simply replacing the Gaussian errors with DPM error terms but leaving the model equations unchanged, we obtain class size effects of -0.103 and -0.108 (with hyperparameter setting "DPM2"). Hence, we find virtually no difference between 4th and 5th graders and estimates close to the 2SLS estimate for 4th graders.
As a robustness check for validity of the instrument, Angrist & Lavy (1999) add linear, quadratic and piecewise linear effects of enrollment to the equations and find that this has quite an impact on the estimated coefficients for class size (ranging between -0.074 and -0.147 and between -0.186 and -0.275 for fourth and fifth graders, respectively). That is, inclusion of *enrol* and the functional form of its effect (which is roughly approximated by a few parametric specifications) affects the estimated class size effect. Furthermore, a violation of the linearity assumption on the class size effect cannot be ruled out and there may be a positive effect for small classes which vanishes for larger classes above some kind of threshold. This would correspond to a nonlinear effect, which could not properly be identified by a simple linear model. To address these issues, we relax the assumption of linear effects and extend the model of Angrist & Lavy (1999) to the following specification

$$\begin{aligned} tscore_{ji} &= \gamma_{20} + f_{21}(csize_{ji}) + f_{22}(disadv_{ji}) + f_{23}(enrol_j) + \varepsilon_{2ji}, && (4.6)\\ csize_{ji} &= \gamma_{10} + \gamma_{11} pcsize_{ji} + f_{12}(disadv_{ji}) + f_{13}(enrol_j) + \varepsilon_{1ji}. && (4.7) \end{aligned}$$

Note that inclusion of random school effects $\nu_{rj} \sim \mathrm{N}(0, \sigma^2_{\nu_r})$ with inverse gamma priors on the variance parameters $\sigma^2_{\nu_r} \sim \mathrm{IG}(a_{\sigma_{\nu r}}, b_{\sigma_{\nu r}})$, $r = 1, 2$ in both equations capturing within-school correlations of class average scores did not change the results substantively but basically only increased the widths of the confidence bands slightly and are therefore not discussed further. Also note that within-school correlations will be generally positive and thus will increase confidence band width (given point estimates do not change) such that given confidence bands will not underestimate estimation precision.
Figure 4.3 shows estimated smooth effects for 4th graders (top panels) and 5th graders (bottom panels) in Equation (4.6) (solid black lines) jointly with 95% pointwise credible intervals (gray areas) and 95% simultaneous credible bands (areas between black dashed curves). On the left hand side, class size effects together with 2SLS estimates in the model excluding *enrol* (gray solid line) and including a linear (gray dashed line) and quadratic effect (gray dotted line) of *enrol* are given. All results are based on hyperparameter specification "DPM2", results with "DPM1" were very similar. Recall that curves are centered around zero (with respect to the covariate values) to ensure identifiability.
Regarding 4th grade students, no significant class size effect is found. This does not

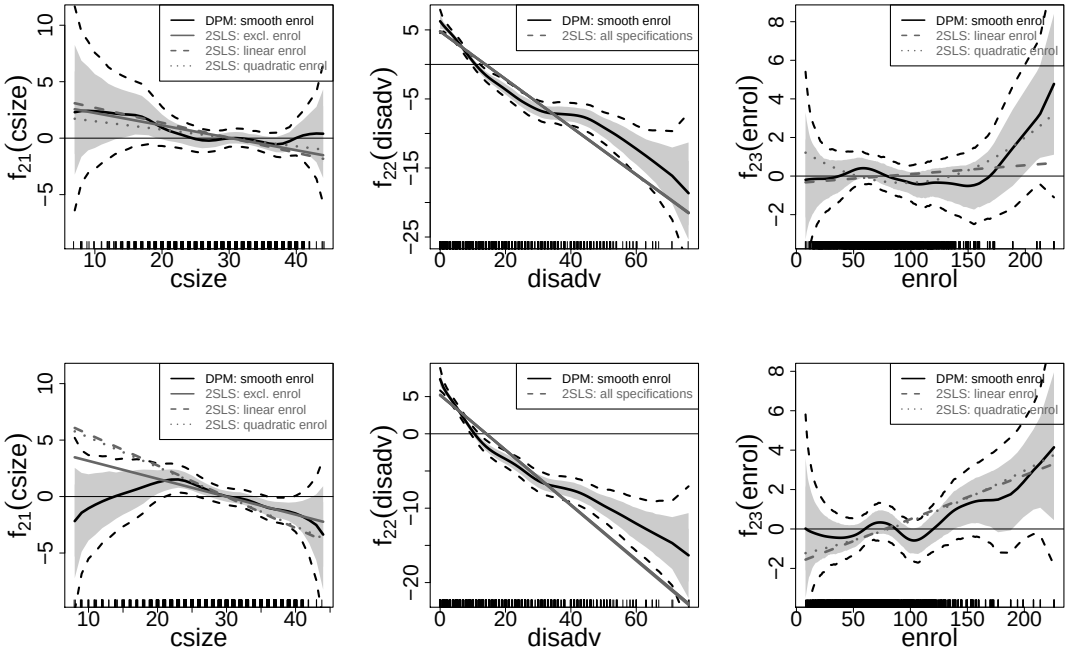

Figure 4.3: Estimated effects for 4th (top) and 5th grade (bottom) students. Solid black lines show smooth curves in Equation (4.6) with 95% pointwise (gray areas) and simultaneous (areas between dashed lines) credible bands. 2SLS results for different parametric specifications of enrolment are given by gray lines.

mean, however, that there is none, the data (and instrument) might just be not informative enough. Note that using 2SLS, the functional form specification of the enrolment effect (not included, linear, quadratic or piecewise linear) has a relatively strong impact on the class size coefficient. In contrast, using the nonparametric DPM approach, inclusion of a smooth effect of enrolment barely influenced the class size effect and therefore results for the model without enrolment are omitted. Revealed by the simultaneous credible bands, estimation uncertainty is excessively high particular for class sizes smaller than 20 casting interpretability of point estimates into doubt. If, however, one is willing to do so, we find indeed a negative relationship between class size and student performance for small class sizes (less then 25 students) and no association as soon as this "threshold" is exceeded.

For fifth grade students, again estimation uncertainty is too high to draw reliable conclu-

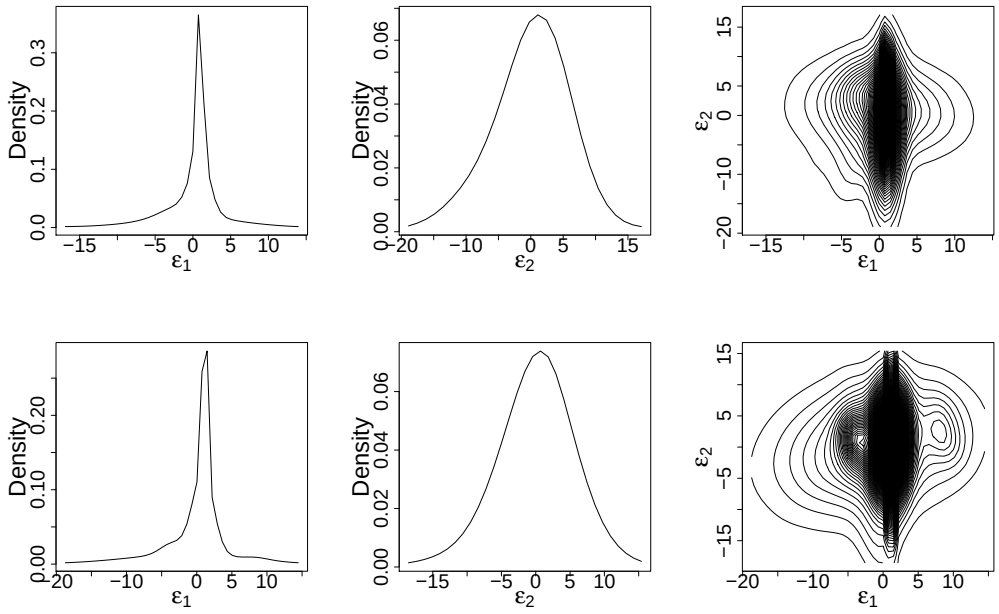

Figure 4.4: Estimated marginal and joint error densities for 4th (top) and 5th grade (bottom) students.

sions on the impact of *csize* on students performance and its functional form. However, note that we find a significant deviation from the linear 2SLS fit. Also note that pointwise intervals (gray areas in Figure 4.3) clearly understate the uncertainty (for the whole curve) and interpreting them would lead to the conclusion of a significant effect, which is however not justified.

For both grades, the estimated curves $\hat{f}_{22}(disadv)$ (see Figure 4.3 middle plots) significantly deviate from the linear estimates obtained from 2SLS (gray straight lines). Such a misspecification of the functional form of the effect of a control variable can of course also affect the estimated class size effect. The smooth effects of enrolment are highly nonlinear but not significant for both grades.

In Figure 4.4, error densities are given which are clearly nonnormal. In particular, the error density for the first equation has a distinct peak while both densities show some slight indication of asymmetry.

It is also interesting to note that using the proposed approach we obtain $\hat{\gamma}_{11} \approx 0.99$ in the first stage equation which is very close to the theoretically expected coefficient

equal to 1. Angrist & Lavy (1999) obtained coefficient estimates of 0.772 and 0.670 and of 0.702 and 0.542 for fourth and fifth graders, respectively, and depending on whether (a linear effect of) *enrol* was included or not. Thus, they obtain substantially smaller coefficients than expected leading to different bias correction. Differences most likely occur due to different handling of outliers in 2SLS and the Bayesian model based on the DPM prior.
Finally, note that Horowitz (2011) analyzed the same data with a bivariate smooth function of *csize* and *disadv*. They also find no significant class size effect (though only reporting results for $disadv = 1.5$).

4.6 Conclusion

We presented a flexible, nonparametric approach for models with one endogenous regressor. The advantages include the availability of simultaneous credible intervals, which naturally incorporate the variability of estimation of the instrumental variable equation. They also work well in small samples and are not only asymptotically correct. We do not rely on a normality assumption such that violations of bivariate normality will not affect estimates and more efficient interval estimates are provided. In our simulation study, we show that the approach based on the DPM is quite robust in case of outliers making the Bayesian approach advantageous even in the parametric context, where although 2SLS methods are consistent they are sensitive to outliers in finite samples. Further, the smoothing parameters controlling the wiggliness of the curves are estimated from the data. In contrast to two-step frequentist approaches we do not have to worry about the difficult smoothing parameter selection for the control function. Our method can also easily be extended to incorporate additive spatial effects based on Gaussian Markov random field priors, smooth interaction terms and varying coefficients based on the framework of structured additive regression (Fahrmeir, Kneib & Lang, 2004).
In our application, we found that without imposing linearity on effects, no reliable conclusions on the relationship between class sizes and student performance can be drawn. Interesting questions for future research include the incorporation of discrete endogenous variables and binary/categorical outcomes of interest as well as nonparametric sample selection models adjusting the error density estimation in Wiesenfarth & Kneib (2010). Our results can also be used for seemingly unrelated regression (SUR) extending Lang, Adebayo, Fahrmeir & Steiner (2003).
The approach is implemented in an R package aiming at providing the method to a wide range of practitioners, see Section 5.2 for details.

5 Software

5.1 Package AdaptFitOS

The approach proposed in Chapter 2 is implemented in a comprehensive R package `AdaptFitOS` based on package `AdaptFit` (Krivobokova, 2009) (which is itself based on the `SemiPar` package (Wand, 2010)). Particular differences to `AdaptFit` include the availability of simultaneous confidence bands and B-spline basis functions and different functionality of the `plot()` function. However, random effects, autocorrelations and interaction surfaces as well as non-Gaussian responses are only limitedly supported. Note that in contrast to `AdaptFit` and `SemiPar`, estimated curves are centered to have zero mean and unlike `SemiPar`, categorical covariates are automatically detected. The package comes with a comprehensive documentation and examples which can be accessed via the common R help system (i.e. using `?AdaptFitOS-package` for instance).
Generally, to fit the model the core fitting function `asp2()` is used which fits semiparametric regression models using the mixed model representation of penalized splines with possibly spatially adaptive penalties. Using the resulting object, fitted curves or their derivatives can be plotted using `plot()`. The usual information on parametric effects as well as results from the specification test proposed in 2.5 can be printed using `summary()`. In the remainder, the functionality of the package is illustrated by the analysis of undernutrition in Kenya. For more details on the individual functions and additional capabilities consult the help documentation. We suppose that the data (available at `http://www.measuredhs.com`) with appropriate labeling of the variables is loaded and attached.

5.1.1 Fitting a Non-adaptive Model

First, we consider the basic model (2.5) in Section 2.3 and display the corresponding plots (Figure 2.3).
At first, vectors of knots for nonparametric effects have to be created. Since we will use B-splines, actually only the length of the vectors will be used as information on the number of knots. Vectors of knots can be created by using

```
> kn.age     = default.knots(age, 40)
> kn.bmi     = default.knots(bmi, 30)
> kn.mheight = default.knots(mheight, 30)
```

In order to fit the model, function `asp2()` is used with an `aspFormula` object specifying the model formula as main argument. Thereby, nonparametric effects are specified by `f(.)`. Within the parenthesis, the covariate, the basis (`"os"` for B-splines as the recommended default), the degree (and possibly the penalty order in case of B-splines), a vector of knots and a logical argument `adap` specifying whether locally adaptive smoothing should be applied, are given. For the first model we use nonadaptive smoothing parameters and B-splines with degree $p = 5$ and penalty order $q = 3$ for all nonparametric curves

```
> fit1= asp2(
   Z ~ f(age, basis="os", degree=c(5,3), knots=kn.age, adap=FALSE)
    + f(bmi, basis="os", degree=c(5,3), knots=kn.bmi, adap=FALSE)
    + f(mheight, basis="os", degree=c(5,3), knots=kn.mheight, adap=FALSE)
    + yearsofedu + rural + female + region)
```

If no basis and knots are given, B-spline bases with $p = 3$, $q = 2$ and the number of knots according to a rule of thumb are used. Note that `region` is a categorical variable. In Figure 2.3, we were not interested in confidence bands (`bands=FALSE`), but wished to display the partial residuals (`residuals=TRUE`). This can be accomplished by specifying

```
> plot(fit1, bands=FALSE, residuals=T, residuals.col=grey(0.4), pages=1)
```

By specifying `pages=1`, all components are plotted in one window, use `pages=0` (default) in order to leave all graphics settings as they are.

5.1.2 Fitting a Model with Locally Adaptive Smoothing Parameters and Heteroscedastic Errors

The model in Section 2.7 is fitted in two steps. First, we refit the previous model with locally-adaptive smoothing parameter for $f_1(age)$. This is accomplished by

```
> kn.lambda= default.knots(kn.age, 5)
> fit2= asp2(
   Z ~ f(age, basis="os", degree=c(5,3), knots=kn.age,
         var.basis="tps", var.degree=3, var.knots=kn.lambda, adap=TRUE)
    + f(bmi, basis="os", degree=c(5,3), knots=kn.bmi, adap=FALSE)
    + f(mheight, basis="os", degree=c(5,3), knots=kn.mheight, adap=FALSE)
    + yearsofedu + rural + female + region, niter.var=300)
```

Here, `kn.lambda` is the vector of knots τ_{w_1} of length $k_w = 5$ for the smoothing parameter function for age, which is modeled with radial basis functions (`var.basis="tps"`) of degree 3 (`var.degree=3`).
In the second step, the varying residual variance (with respect to bmi) is estimated with cubic B-splines and $k_v = 5$ knots using

```
> fit2B= aspHetero(fit2, xx=bmi, basis="os", degree=c(3,2), nknots=5)
```

We can now plot all fitted curves jointly with heteroscedasticity adjusted simultaneous confidence bands (Figure 2.4) using

```
> plot(fit2B, level=0.95,
       xlab= list("age (in months)",
                  expression("bmi (in "* kg/m^2*")"), "mheight (in mm)"))
```

Here, labels for the x-axes were specified in a list of length equal to the number of smooth curves in the model. Otherwise, axes are labeled in an automatic fashion. Of course, the layout can be adjusted with additional arguments such that confidence bands are shaded or lines are thicker.
Since in large data sets, estimation of $\widehat{\text{Var}}(\hat{f})$ can be memory intensive and take a couple of minutes, we can also first create an `scbm` object using `scbM()` which can then be plotted much faster using the same arguments as above.

```
> scb2B= scbM(fit2B)
> plot(scb2B)
```

The varying residual variance can be extracted with the auxiliary function `sigma()` which can be used for plotting purposes (Figure 2.5(b))

```
> plot(sort(bmi), sigma(fit2B)[order(bmi)], type="l",
       xlab=expression("bmi (in "* kg/m^2*")"),
       ylab=expression(sigma(bmi)))
```

Estimated derivatives (as in Figure 2.6) are plotted by specifying the derivative order in the plot function

```
> plot(fit2B, select=1, drv=1)
```

Since we are only interested in the first derivative of the first function, we specified `select=1`.
Coefficient estimates with corresponding standard errors and p-values are printed using the common `summary()` function

```
> summary(scb2B)
```

Using additional logical arguments `test1` and `test2`, tests for no-effect and the nonparametric specification test proposed in Section 2.5 can be printed (with significance level given by argument `signif`), respectively. The test for no-effect corresponds to checking whether the zero line can be drawn inside the simultaneous confidence band around the nonparametrically estimated curve and – in contrast to the specification test – does not depend on the used penalty order q. Its test statistic is defined as

$$T_j^0 = \sup_{x \in [0,1]} \left(|\hat{f}_j(x)| / \sqrt{\mathrm{Var}\{\hat{f}_j(x)\}} \right).$$

Rejection of H_0 takes place if $T_j^0 > c_{m,j}$. Note that this test coincides with the nonparametric specification test in Section 2.5 for $q = 1$. Thus, for $q = 1$ its power will be close to the RLR test, but for $q > 1$ improved power can be expected due to stronger smoothness assumptions imposed.

```
> summary(scb2B, test1=TRUE, test2=TRUE, signif=0.05)

Summary for linear components:

                        coef       se    ratio p-value
intercept           -1.38000 0.091430 -15.0900  0.0000
yearsofedu           0.04063 0.005995   6.7780  0.0000
rural               -0.10030 0.060440  -1.6590  0.0972
female               0.19740 0.040550   4.8680  0.0000
regioncentral       -0.13160 0.101300  -1.3000  0.1936
regioncoast         -0.07125 0.101800  -0.6999  0.4840
regioneastern       -0.16290 0.105900  -1.5380  0.1240
regionnyanza        -0.12930 0.101500  -1.2740  0.2026
regionrift valley   -0.15140 0.097520  -1.5530  0.1205
regionwestern       -0.24750 0.099190  -2.4950  0.0126
regionnorth eastern  0.54690 0.126700   4.3160  0.0000


Summary for non-linear components:

         basis deg pen  adap  knots   |  tstat crit(0.05%) pval
f(age)      os   5   3  TRUE     40   | 20.001       3.101    0
f(bmi)      os   5   3 FALSE     30   |  8.217       2.968    0
f(mheight)  os   5   3 FALSE     30   | 14.519       2.900    0
```

```
Test for a polynomial of degree...:

            degree  adap  tstat crit(0.05%)  pval
f(age)           2  TRUE 11.044       3.176 0.000
f(bmi)           2 FALSE  2.415       3.090 0.339
f(mheight)       2 FALSE  1.279       3.167 1.000
```

Results for parametric effects are given in the first block, tests for no-effect of nonparametrically estimated effects in the second and results using the nonparametric specification test are given in the last block. Since we used $q = 3$ for all curves, we tested for deviations from quadratic fits in all cases.
To assess test results of the nonparametric specification test in case of large data sets, function `scbTest()` can also be used which can be convenient in combination with argument `select` when results are only needed for a subset of the additive components. The resulting object can also be plotted using `plot()`, returning the estimated deviation from the function under the null hypothesis with corresponding simultaneous confidence bands which can help to get an idea about the intuition of the test.
Recall that random effects, autocorrelations and interaction surfaces as well as non-Gaussian responses are not supported by `asp2()` since simultaneous confidence bands are not yet available for these cases. The package contains, however, two functions `aspOS()` and `spmOS()` extending functions `asp()` (package `AdaptFit`) and `spm()` (package `Semipar`) to incorporate B-splines with penalty with respect to an integrated squared derivative. These can be used for instance to conduct the RLRT-test using the `RLRsim` package as done in Section 2.5.

5.2 Package bayesIV

Both the geoadditive sample selection model (Chapter 3) and the model with continuous righthand side endogenous variable (Chapter 4) are implemented in an R package, called `bayesIV`. Full description and examples are provided in its documentation which can be accessed via the usual R help system.
There are four estimation routines: `am()` for estimation of (single equation) geoadditive regression models, `ssm()` for estimation of geoadditive sample selection models and `bayesIV()` and `bayesIVgauss()` for estimation of (geo-)additive instrumental variable regression models with DPM and normal-inverse Wishart priors on the error densities, respectively. Plotting and diagnostic capabilities are shared by all estimation routines.

In the next subsection, function `am()` for estimation of univariate (geo-)additive regression models is briefly introduced. In Sections 5.2.2 and 5.2.3 computation of a Bayesian geoadditive sample selection model and of a nonparametric instrumental variable regression problem are explained via the analysis of relief supply in Pakistan and the relationship between class size and student performance, respectively.

5.2.1 Bayesian Geoadditive Regression Models

Function `am()` provides an implementation for estimation of univariate Bayesian (geo-)additive regression models as described in Lang & Brezger (2004). Componentwise simultaneous credible bands as proposed by Crainiceanu, Ruppert, Carroll, Joshi & Goodner (2007), Besag, Green, Higdon & Mengersen (1995) and Krivobokova, Kneib & Claeskens (2010) can be added to the fits (adjusting function `scbB()` of package `ConfBands`). Plotting capabilities of package `BayesX` are exploited for the drawing of spatial effects. It was mainly implemented to provide a method for comparison of single equation estimates (without sample selection or endogeneity bias correction) with the methods proposed in Chapters 3 and 4.
Its usage follows the basic syntax of all estimation routines in the package

```
> am(form, random = list(NULL), data,
        numKnots = 20, degree = 3, rw.order = 2,
        numBurnIn = 2000, numSamples = 20000, thin = 20)
```

The model formula is given in `form`. Smooth functions are specified by `s(.)`, varying-coefficients by `s(., by=.)` and spatial effects by `sp(.)`. Smooth curves are modeled with B-splines of degree `degree` and random walk penalty of order `rw.order`. In argument `random`, cluster IDs can be given for inclusion of random intercepts. `numBurnIn` describes how many iterations are discarded for burn-in. `numSamples` is the number of subsequent iterations from which only every `thin`-th iteration is used for inference. See the documentation on how to specify hyperparameters for prior distributions.
A `summary()` function provides information on parametric estimates. Resulting smooth curves can be plotted using `plot()` and spatial effects can be plotted using `plotmaps()`. For diagnostics, `plotacf()` and `plotpaths()` provide tools to plot the posterior sample autocorrelations and sampling paths, respectively. `plotpost()` provides posterior densities of parametric effects. Further plotting capabilities are described in the following.

5.2.2 Bayesian Geoadditive Sample Selection Models

Sample selection models are estimated with function `ssm()`. The Gibbs sampling scheme involved is based on the joint posterior distribution marginalized over unobserved ob-

servations as described in Chapter 3. Its usage is close to the `am()` function replacing the `form` argument by two formula objects `selection` and `outcome` for specification of selection and outcome equation, respectively. Thus, smooth curves, varying coefficients and spatial effects are specified in the same way as above. The dependent variable in the selection equation determines whether the dependent variable in the outcome equation is observed. It is assumed to be binary with 0 when the response of the outcome equation is unobserved and 1 if the response variable of the outcome equation is observed. The response variable of the outcome equation is assumed to be Gaussian when it is observed. In the following, the analysis of relief supply in Pakistan (commodity type "food, kitchen supplies & water") is used to illustrate the functionality of function `ssm()`. The model equations with variable labels as given in Section 3.5.2 are given by

```
> sel = selectfood ~ height + lnheli + acc +
           s(dist) + s(t, by=rug) + s(t, by=MMI) + s(t, by=pop) + sp(s)
> out = lnfood ~ height + lnheli + acc +
           s(dist) + s(t, by=rug) + s(t, by=MMI) + s(t, by=pop) + sp(s)
```

Then, the model can be fitted using

```
> food= ssm(selection=sel, outcome=out, numKnots=30, rw.order=2,
  numBurnIn=20000, numSamples=80000, thin=80,
  graph.sel= "neighborstruct")
```

Since spatial effects of regional variable s are present, a `graph` object (see package `BayesX`) – called "neighborstruct" here – containing information on neighborhood structures is given. Regional variables on different administrative levels in the two equations are also supported (in which case an additional `graph.out` argument has to be specified). It is recommended to choose large numbers of `numBurnIn`, `numSamples` and `thin` and to check sample autocorrelations (using `plotacf()`) due to a tendency of high sample autocorrelations of parameters in the selection equation. By default, uninformative priors are used as described in Chapter 3. For adjustments of the prior settings, consult the manual.
Besides information on the parametric coefficients, the `summary()` function additionally returns information on (co-)variance estimates and the DIC (Deviance Information Criterion) for model selection.

```
> summary(food)

Formulas:
 Selection equation: selectfood ~ height + lnheli + acc + s(dist) + s(t, by=rug)
```

```
 + s(t, by=MMI) + s(t, by=pop) + sp(s)
 Outcome equation:   lnfood ~ height + lnheli + acc + s(dist) + s(t, by=rug)
 + s(t, by=MMI) + s(t, by=pop) + sp(s)

17313 observations (16469 censored and 844 observed)
Number of samples discarded (burn-in): 20000;
Number of Samples used: 1000 of 80000 Samples (thinning=80)

Parametric coefficients:

Selection equation:
             Estimate Std.Dev. 2.5%quant.   Median 97.5%quant. p-value(2sided)
(Intercept) -10.0451    2.1097   -13.9828 -10.0279     -5.8167           0.000
height        0.0014    0.0005     0.0005   0.0013      0.0024           0.000
lnheli        0.7169    0.2909     0.0926   0.7109      1.2530           0.022
acc           0.0759    0.0707    -0.0617   0.0768      0.2152           0.302

Outcome equation:
           Estimate Std.Dev. 2.5%quant.  Median 97.5%quant. p-value(2sided)
(Intercept) 35.3435  6.57887    21.5883 35.7198     47.1118           0.000
height      -0.0004  0.00061    -0.0016 -0.0004      0.0008           0.492
lnheli      -1.1028  0.97514    -2.7650 -1.1924      0.9345           0.290
acc         -0.1831  0.17507    -0.5200 -0.1762      0.1626           0.286


SIGMA:
         [,1]     [,2]
[1,]  1.00000 -1.62918
[2,] -1.62918  3.18026

Quantiles of the covariance components and correlation between disturbances:
                 Mean      2.5       5%      50%      95%    97.5%
cov(sel,out) -1.62918 -1.90950 -1.86462 -1.63446 -1.38252 -1.30204
Var(outcome)  3.18026  2.43474  2.56105  3.18884  3.88065  4.01938
Correlation  -0.91356 -0.95394 -0.94936 -0.91533 -0.85606 -0.83437

DIC: 33510.69
```

Smooth curves as in Figure 3.3 are displayed using `plot(food)`. Note that simultaneous credible bands as described in Section 4.3.4 are also provided for the sample selection model (which has not yet been discussed in Chapter 3). The spatial effects in the outcome equation (Figure 3.5 right column) are plotted using

```
> plotmaps(food, equation=2, map=mapobject)
```

where `mapobject` is an object containing the required boundary information (as obtained by a call to `read.bnd` of package `BayesX`).

5.2.3 Bayesian Nonparametric Instrumental Variable Regression

Function `bayesIV()` provides an implementation of the DPM approach proposed in Chapter 4. An implementation of the Bayesian nonparametric instrumental variable regression model assuming bivariate normal errors as standard in the literature is provided by function `bayesIVgauss()` with essentially the same syntax.
In the following the functionality is illustrated via the analysis of the effect of class size on student performance given in Section 4.5. The data is available on `http://emlab.berkeley.edu/users/card/data_sets.html` and is assumed to be loaded as object `data` with appropriate variable labeling.
The model equations are given by

```
> first = csize ~ pcsize + s(disadv) + s(enrol)
> second = tscore ~ s(csize) + s(disadv) + s(enrol)
```

The model is fitted (with hyperparameter specification "DPM2") using

```
> bivDPM2= bayesIV(first=first, second=second,
                   numBurnIn=5000, numSamples=40000, thin=40,
                   priorDPM= list(
                    # fixed parameters in G0
                      s.Sigma=3, S.Sigma=0.2*diag(2),
                    # hyperprior on tau
                      a.Sigma=1, b.Sigma=100,
                    # gamma prior on alpha
                      a.alpha=2, b.alpha=2 ))
```

Thus, formula objects `selection` and `outcome` in function `ssm()` are replaced by formula objects `first` and `second` specifying "first stage" (with right hand side endogenous variable y_1 as response) and "second stage" equation (the equation of primary interest), respectively. The model assuming bivariate normal errors is fitted in the same way using function `bayesIVgauss()` (with hyperparameters `s.Sigma` and `S.Sigma` only). The `priorDPM` argument is the default and could thus be omitted. Hyperparameter setting "DPM1" is obtained by adding `tau.Sigma=.036` to the list (discarding the parameters `a.Sigma` and `b.Sigma` for the gamma prior on τ_Σ). See the documentation for further possibilities of hyperparameter adjustments.
By default, adjusted routines of package `DPpackage` are exploited for density estimation. Alternatively, the sampling algorithm of Conley, Hansen, McCulloch & Rossi (2008) can be used (with different Gibbs sampler and discrete hyperprior on the concentration parameter α) by specifying `density.method="conley"`.

The `summary()` function prints information on parametric effects such as coefficient estimates, credible intervals and posterior probabilities. For grade 4, we obtain (slightly shortened for the sake of brevity)

```
> summary(bivDPM2)

Formulas:
 First equation: classize ~ func1 + s(tipuach) + s(c_size)
 Second equation:   avgverb ~ s(classize) + s(tipuach) + s(c_size)

2049 observations
Number of samples discarded (burn-in): 2000;
Number of Samples used: 1000 of 40000 Samples (thinning=40)

Distribution of error terms flexibly estimated

Parametric coefficients:

First equation:
            Estimate    mgcv Std.Dev. 2.5%quant.  Median 97.5%quant. p-value(2sided)
(Intercept)  -0.2977 15.4531   0.2362    -0.7452 -0.3031      0.1802           0.198
pcsize        0.9837  0.4769   0.0073     0.9686  0.9838      0.9975           0.000

Second equation:
            Estimate    mgcv Std.Dev. 2.5%quant.  Median 97.5%quant. p-value(2sided)
(Intercept)   72.484 72.4894   0.1271    72.2426 72.4822     72.7202               0
```

All smooth curves in the model including 95% simultaneous and pointwise credible bands (Figure 4.3) can be plotted by using

```
> plot(bivDPM2)
```

In order to plot only the curves in one equation, use additional argument `equation`. If only a single curve is to be fitted the covariate name can be additionally given, e.g. in order to plot the endogenous class size effect only, we type

```
> plot(bivDPM2, equation=2, covariate="csize").
```

We plot the marginal and joint error densities (Figure 4.4) using

```
> density(bivDPM2)
```

For layout adjustments of the plot functions we refer to the help documentation. Note that spatial effects (based on Markov random field priors as described in Section 3.2.4) are also provided for instrumental variable regression models and can be plotted using `plotmaps()`.

References

Albert, J. H. and Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American Statistical Association* **88**, 669–679.

Anand, S. and Kanbur, S. (1993). Inequality and development a critique. *Journal of Development economics 41*(1), 19–43.

Angrist, J. and Lavy, V. (1999). Using Maimonides' Rule to Estimate The Effect of Class Size on Scholastic Achievement. *Quarterly journal of economics 114*(2), 533–575.

Antoniak, C. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The Annals of Statistics 2*(6), 1152–1174.

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 semiparametric regression approach. In T. Kneib & G. Tutz (Eds.), *Statistical Modelling and Regression Structures*, pp. 155–179. Physica-Verlag HD.

Benini, A., Conley, C., Dittemore, B., and Waksman, Z. (2006). Survivor Needs or Logistical Convenience? Factors shaping decisions to deliver relief to earthquake-affected communities, Pakistan 2005–06. Navigating post-conflict environments series, Vietnam Veterans of America Foundation/Information Management and Mine Action Programs (VVAF/iMMAP), Washington, DC.

Benini, A., Conley, C., Dittemore, B., and Waksman, Z. (2009). Survivor needs or logistical convinience? Factors shaping decisions to deliver relief to eartquake-affected communities, Pakistan 2005-06. *Disasters* **33**, 110–131.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995). Bayesian computation and stochastic systems. *Statistical Science 10*(1), 3–41.

Bickel, P. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *The Annals of Statistics 1*(6), 1071–1095.

Blackwell, D. and MacQueen, J. (1973). Ferguson distributions via pólya urn schemes. *The Annals of Statistics 1*(2), 353–355.

Blundell, R. and Powell, J. (2003). Endogeneity in nonparametric and semiparametric regression models. In M. Dewatripont, L. Hansen, & S. Turnovsky (Eds.), *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, Volume 2, pp. 294–311. Cambridge University Press.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. In L. Christofides & R. Swindinsky (Eds.), *Aspects of Labor Market Behavior: Essays in Honor of John Vanderkamp*, pp. 201–222. University of Toronto Press.

Central Bureau of Statistics (CBS) Kenya, Ministry of Health (MOH) Kenya, and ORC Macro (2004). *Kenya Demographic and Health Survey 2003*. Calverton, Maryland: CBS, MOH, and ORC Macro.

Chao, J. and Phillips, P. (1998). Posterior distributions in limited information analysis of the simultaneous equations model using the jeffreys prior. *Journal of Econometrics 87*(1), 49–86.

Chib, S. and Greenberg, E. (2007). Semiparametric Modeling and Estimation of Instrumental Variable Models. *Journal of Computational and Graphical Statistics 16*(1), 86–114.

Chib, S. and Greenberg, E. (2010). Additive cubic spline regression with dirichlet process mixture errors. *Journal of Econometrics 156*(2), 322–336.

Chib, S., Greenberg, E., and Jeliazkov, I. (2009). Estimation of semiparametric models in the presence of endogeneity and sample selection. *Journal of Computational and Graphical Statistics 18*(2), 321–348.

Claeskens, G., Krivobokova, T., and Opsomer, J. (2009). Asymptotic properties of penalized spline estimators. *Biometrika 96*(3), 529–544.

Claeskens, G. and Van Keilegom, I. (2003). Bootstrap confidence bands for regression curves and their derivatives. *The Annals of Statistics 31*(6), 1852–1884.

Conley, T., Hansen, C., McCulloch, R., and Rossi, P. (2008). A semi-parametric Bayesian approach to the instrumental variable problem. *Journal of Econometrics 144*(1), 276–305.

Crainiceanu, C., Ruppert, D., Claeskens, G., and Wand, M. (2005). Exact likelihood ratio tests for penalised splines. *Biometrika 92*(1), 91.

Crainiceanu, C. M., Ruppert, D., Carroll, R. J., Joshi, A., and Goodner, B. (2007). Spatially adaptive bayesian penalized splines with heteroscedastic errors. *Journal of Computational and Graphical Statistics 16*(2), 265–288.

Darolles, S., Fan, Y., Florens, J., and Renault, E. (2011). Nonparametric instrumental regression. *Econometrica 79*(5), 1541–1565.

Das, M., Newey, W., and Vella, F. (2003). Nonparametric estimation of sample selection models. *Review of Economic Studies 70*(1), 33–58.

Dasgupta, S., Laplante, B., Wang, H., and Wheeler, D. (2002). Confronting the environmental kuznets curve. *The Journal of Economic Perspectives 16*(1), 147–168.

Durban, M. and Currie, I. (2003). A note on P-spline additive models with correlated errors. *Computational Statistics 18*(2), 251–262.

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

Escobar, M. and West, M. (1995). Bayesian Density Estimation and Inference Using Mixtures. *Journal of the American Statistical Association 90*(430), 577–588.

Fahrmeir, L., Kneib, T., and Lang, S. (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica 14*(3), 731–761.

Fan, J. and Gijbels, I. (1996). *Local polynomial modelling and its applications*. London: Chapman & Hall.

Harbaugh, W., Levinson, A., and Wilson, D. (2002). Reexamining the empirical evidence for an environmental kuznets curve. *Review of Economics and Statistics 84*(3), 541–551.

Härdle, W., Huet, S., Mammen, E., and Sperlich, S. (2004). Bootstrap inference in semiparametric generalized additive models. *Econometric Theory 20*(2), 265–300.

Härdle, W., Sperlich, S., and Spokoiny, V. (2001). Structural tests in additive regression. *Journal of the American Statistical Association 96*(456), 1333–1347.

van Hasselt, M. (2005). Bayesian sampling algorithms for the sample selection and two-part models. *Computing in Economics and Finance 2005* **241**.

Hastie, T. and Tibshirani, R. (1990). *Generalized additive models*. London and New York: Chapman and Hall.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society, Series B* **55**, 757–796.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Henningsen, A. and Toomet, O. (2008). *sampleSelection: Sample Selection Models*. R package version 0.5-5.

Hoogerheide, L., Kaashoek, J., and Van Dijk, H. (2007). On the shape of posterior densities and credible sets in instrumental variable regression models with reduced rank: an application of flexible sampling methods using neural networks. *Journal of Econometrics 139*(1), 154–180.

Horowitz, J. and Lee, S. (2009). Uniform confidence bands for functions estimated nonparametrically with instrumental variables. *CeMMAP working papers*.

Horowitz, J. L. (2011). Applied nonparametric instrumental variables estimation. *Econometrica 79*(2), 347–394.

Horton, S., Alderman, H., and Rivera, J. (2009). Hunger and malnutrition. In B. Lomborg (Ed.), *Global Crises, Global Solutions*. Cambridge University Press, Cambridge, 2nd ed.

Ishwaran, H. and James, L. (2002). Approximate dirichlet process computing in finite normal mixtures. *Journal of Computational and Graphical Statistics 11*(3), 508–532.

Jara, A., Hanson, T., Quintana, F., Müller, P., and Rosner, G. (2011). DPpackage: Bayesian semi- and nonparametric modeling in R. *Journal of Statistical Software 40*(5), 1–30.

Johansen, S. and Johnstone, I. (1990). Hotelling's theorem on the volume of tubes: some illustrations in simultaneous inference and data analysis. *The Annals of Statistics 18*(2), 652–684.

Kabubo-Mariara, J., Ndenge, G., and Mwabu, D. (2009). Determinants of children's nutritional status in Kenya: Evidence from demographic and health surveys. *Journal of African Economies 18*(3), 363.

Kai, L. (1998). Bayesian inference in a simultaneous equation model with limited dependent variables. *Journal of Econometrics 85*(2), 387–400.

Kandala, N., Fahrmeir, L., Klasen, S., and Priebe, J. (2009). Geo-additive models of childhood undernutrition in three sub-Saharan African countries. *Population, Space and Place 15*(5), 461–473.

Kauermann, G., Krivobokova, T., and Fahrmeir, L. (2009). Some asymptotic results on generalized penalized spline smoothing. *Journal of the Royal Statistical Society, Series B* **71**, 487–503.

Kleibergen, F. and Van Dijk, H. (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory 14*(06), 701–743.

Kleibergen, F. and Zivot, E. (2003). Bayesian and classical approaches to instrumental variable regression. *Journal of Econometrics 114*(1), 29–72.

Koop, G., Poirier, D., and Tobias, J. (2005). Semiparametric Bayesian inference in multiple equation models. *Journal of Applied Econometrics 20*(6), 723–747.

Krivobokova, T. (2009). *AdaptFit: Adaptive Semiparametic Regression*. R package version 0.2-2.

Krivobokova, T., Crainiceanu, C., and Kauermann, G. (2008). Fast adaptive penalized splines. *Journal of Computational and Graphical Statistics 17*(1), 1–20.

Krivobokova, T., Kneib, T., and Claeskens, G. (2010). Simultaneous confidence bands for penalized spline estimators. *Journal of the American Statistical Association 105*(490), 852–863.

Kuznets, S. (1955). Economic growth and income inequality. *The American Economic Review 45*(1), 1–28.

Lang, S., Adebayo, S. B., Fahrmeir, L., and Steiner, W. J. (2003). Bayesian geoadditive seemingly unrelated regression. *Computational Statistics 18*(2), 263–292.

Lang, S. and Brezger, A. (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*(1), 183–212.

Lee, L. (2000). Self-Selection. In B. Baltagi (Ed.), *A Companion to Theoretical Econometrics*, Chapter 18. Blackwell Publishers, Malden, Mass.

Leslie, D., Kohn, R., and Nott, D. (2007). A general approach to heteroscedastic linear regression. *Statistics and Computing 17*(2), 131–146.

Link, W. and Barker, R. (2005). Modeling association among demographic parameters in analysis of open population capture–recapture data. *Biometrics 61*(1), 46–54.

Loader, C. and Sun, J. (1997). Robustness of tube formula based confidence bands. *Journal of Computational and Graphical Statistics 6*(2), 242–250.

Ma, S. and Yang, L. (2011). Spline-backfitted kernel smoothing of partially linear additive model. *Journal of Statistical Planning and Inference 141*(1), 204–219.

Marra, G. and Radice, R. (2011). A flexible instrumental variable approach. *Statistical Modelling 11*(6), 581–603.

Meng, X. and Gelman, A. (1991). A note on bivariate distributions that are conditionally normal. *The American Statistician 45*(2), 125–126.

Min, Y. and Agresti, A. (2002). Modeling nonnegative data with clumping at zero: A survey. *Journal of the Iranian Statistical Society 1*(1-2), 7–33.

Neal, R. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of Computational and Graphical Statistics 9*(2), 249–265.

Neumann, M. H. and Polzehl, J. (1998). Simultaneous bootstrap confidence bands in nonparametric regression. *Journal of Nonparametric Statistics 9*(4), 307–333.

Newey, W. and Powell, J. (2003). Instrumental variable estimation of nonparametric models. *Econometrica 71*(5), 1565–1578.

Newey, W., Powell, J., and Vella, F. (1999). Nonparametric estimation of triangular simultaneous equations models. *Econometrica 67*(3), 565–603.

Nychka, D. (1988). Bayesian confidence intervals for smoothing splines. *Journal of the American Statistical Association 83*(404), 1134–1143.

Omori, Y. (2007). Efficient gibbs sampler for bayesian analysis of a sample selection model. *Statistics & Probability Letters 77*(12), 1300–1311.

Pelletier, D. (1994). The relationship between child anthropometry and mortality in developing countries: implications for policy, programs and future research. *Journal of nutrition* **124**, 2047–2081.

Pinkse, J. (2000). Nonparametric two-step regression estimation when regressors and error are dependent. *Canadian Journal of Statistics 28*(2), 289–300.

Reiss, P. T. and Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society, Series B 71*(2), 505–523.

Robert, C. (1995). Simulation of truncated normal variables. *Statistics and Computing 5*(2), 121–125.

Rue, H. and Held, L. (2005). *Gaussian Markov random fields: Theory and applications*, Volume 104. CRC / Chapman & Hall, London.

Ruppert, D., Wand, M., and Carroll, R. (2003). *Semiparametric Regression*. Cambridge, U.K.: Cambridge University Press.

Scheipl, F., Greven, S., and Küchenhoff, H. (2008). Size and power of tests for a zero random effect variance or polynomial regression in additive and linear mixed models. *Computational Statistics & Data Analysis 52*(7), 3283–3299.

Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* **4**, 639–650.

Severini, T. A. (2000). *Likelihood Methods in Statistics*. Oxford: Oxford University Press.

Sigelman, L. and Zeng, L. (1999). Analyzing censored and sample-selected data with Tobit and Heckit models. *Political Analysis 8*(2), 167.

Song, Q. and Yang, L. (2010). Oracally efficient spline smoothing of nonlinear additive autoregression models with simultaneous confidence band. *Journal of Multivariate Analysis 101*(9), 2008–2025.

Stern, D. (2004). The rise and fall of the environmental kuznets curve. *World development 32*(8), 1419–1439.

Su, L. and Ullah, A. (2008). Local polynomial estimation of nonparametric simultaneous equations models. *Journal of Econometrics 144*(1), 193–218.

Sun, J. (1993). Tail probabilities of the maxima of Gaussian random fields. *The Annals of Probability 21*(1), 34–71.

Sun, J. and Loader, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *The Annals of Statistics 22*(3), 1328–1345.

UNICEF (1998). *The state of the world's children 1998*. Oxford University Press, for UNICEF.

Vella, F. (1998). Estimating models with sample selection bias: A survey. *The Journal of Human Resources 33*(1), 127–169.

Victora, C., de Onis, M., Hallal, P., Blossner, M., and Shrimpton, R. (2010). Worldwide timing of growth faltering: revisiting implications for interventions. *Pediatrics 125*(3).

Wahba, G. (1983). Bayesian confidence intervals for the cross-validated smoothing spline. *The Annals of Statistics 45*(1), 133–150.

Wand, M. (2010). *SemiPar: Semiparametic Regression*. R package version 1.0-3.

Wang, J. and Yang, L. (2009). Efficient and fast spline-backfitted kernel smoothing of additive models. *Annals of the Institute of Statistical Mathematics 61*(3), 663–690.

Wang, L. and Yang, L. (2007). Spline-backfitted kernel smoothing of nonlinear additive autoregression model. *The Annals of Statistics 35*(6), 2474–2503.

WHO (1995). Physical status: The use and interpretation of anthropometry. Technical Report 854, WHO Technical Report Series, Geneva: WHO.

WHO (2006). WHO child growth standards based on length/height, weight and age. *Acta Paediatrica 95*(Supplement 450), 76–85.

Wiesenfarth, M. and Kneib, T. (2010). Bayesian geoadditive sample selection models. *Journal of the Royal Statistical Society: Series C (Applied Statistics) 59*(3), 381–404.

Winship, C. and Mare, R. D. (1992). Models for sample selection bias. *Annual Reviews in Sociology 18*(1), 327–350.

Wood, S. (2006). *Generalized Additive Models*. Chapman and Hall.

Wooldridge, J. (2002). *Econometric analysis of cross section and panel data*. The MIT press.

Yatchew, A. (1998). Nonparametric regression techniques in economics. *Journal of Economic Literature 36*(2), 669–721.