Courant Research Centre

'Poverty, Equity and Growth in Developing and Transition Countries: Statistical Methods and Empirical Analysis'

Georg-August-Universität Göttingen
(founded in 1737)

Discussion Papers

No. 127

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

Manuel Wiesenfarth, Carlos Matías Hisgen, Thomas Kneib, Carmen Cadarso-Suarez

October 2012

Wilhelm-Weber-Str. 2 · 37073 Goettingen · Germany
Phone: +49-(0)551-3914066 · Fax: +49-(0)551-3914059

Email: crc-peg@uni-goettingen.de Web: http://www.uni-goettingen.de/crc-peg

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

Manuel Wiesenfarth[1] Carlos Matías Hisgen[2] Thomas Kneib[3]

Carmen Cadarso-Suarez[4]

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*.

Key words and phrases. Endogeneity; Markov Chain Monte Carlo methods; Simultaneous credible bands.

[1]University of Mannheim, Department of Economics, L7 3-5, 68131 Mannheim, Germany

[2]Department of Economics, Universidad Nacional del Nordeste, Argentina

[3]Department of Statistics and Econometrics, Georg-August-Universität Göttingen, Germany

[4]Department of Biostatistics, School of Medicine, University of Santiago de Compostela, Santiago de Compostela, Spain

1 Introduction

One frequently encountered problem in regression analysis in particular in case of observational data 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. 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 therefore routinely applied in the parametric regression context (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. Thus, in recent years an increasing number of approaches to nonparametric instrumental variable regression has appeared, see Blundell and 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 partly due to a lack of easily available implementations and the need of user assistance for choosing specific tuning parameters. This paper 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: control function approaches and instrumental variable approaches. The *control function approach* (Newey et al., 1999, Pinkse, 2000 and Su and 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

$$y_2 = f_2(y_1) + \varepsilon_2, \qquad y_1 = f_1(z_1) + \varepsilon_1 \tag{1}$$

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

$$\mathrm{E}(y_2|y_1, z_1) = f_2(y_1) + \mathrm{E}(\varepsilon_2|\varepsilon_1, z_1) = f_2(y_1) + \mathrm{E}(\varepsilon_2|\varepsilon_1) = f_2(y_1) + v(\varepsilon_1) \tag{2}$$

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 and 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_1 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 et al., 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 and 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 et al., 2011). Again, also

construction of simultaneous confidence bands is difficult, with Horowitz and Lee (2009) being the first attempt. In the remainder of this paper 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 (see also Kleibergen and Zivot (2003) for an overview over Bayesian parametric methods). All of these assume bivariate normality of the errors $(\varepsilon_1, \varepsilon_2) \sim \mathrm{N}(0, \Sigma)$ (e.g. Chib and Greenberg, 2007, Chib et al., 2009 and Koop et al., 2005). Then, both equations in (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 $\sigma_{12} = \mathrm{Cov}(\varepsilon_{1i}, \varepsilon_{2i})$ and $\sigma_1^2 = \mathrm{Var}(\varepsilon_{1i})$. 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 the observed explanatory variables. 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 $\mathrm{E}(\varepsilon_2|\varepsilon_1)$. 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 et al. (2008) propose the application of a Dirichlet process mixture (DPM) prior (Escobar and 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 and 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 and 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 et al. (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 and 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 paper is organized as follows. In Section 2 the considered model is introduced and prior distributions are discussed. Section 3 describes Bayesian inference including smoothing parameter determination and construction of simultaneous credible bands. In Section 4, small sample properties are explored through simulations and the approach is compared to existing approaches. In Section 5, an application to class size effects on student performance is provided and the paper concludes in Section 6.

2 Additive Simultaneous Equations Model

We consider an additive simultaneous equations model

$$y_{2i} = \gamma_{20} + f_{21}(y_{1i}) + \sum_{\ell=1}^{q_2} x_{2\ell i}\gamma_{2\ell} + \sum_{\ell=1}^{p_2} f_{2,\ell+1}(z_{2\ell i}) + \varepsilon_{2i} \tag{3}$$

$$y_{1i} = \gamma_{10} + \sum_{\ell=1}^{q_1} x_{1\ell i}\gamma_{2\ell} + \sum_{\ell=1}^{q_2} x_{2\ell i}\gamma_{1,q_1+\ell} + \sum_{\ell=1}^{p_1} f_{1\ell}(z_{1\ell i}) + \sum_{\ell=1}^{p_2} f_{1,p_1+\ell}(z_{2\ell i}) + \varepsilon_{1i},\ i = 1,\ldots,n \tag{4}$$

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 and Greenberg (2007), Chib et al. (2009) or Koop et al. (2005)) it implies strong implicit assumptions on the control function as discussed in the introduction. We therefore follow Conley et al. (2008) and employ a Dirichlet process mixture prior (Escobar and West, 1995) for the joint error distribution which basically allows to specify a hyperprior on the space of potential error distributions. Prior choices for all involved parameters are discussed in the following.

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 e.g. Chao and Phillips, 1998, Hoogerheide et al., 2007, Kleibergen and Van Dijk, 1998 and Kleibergen and 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 or to use informative normal priors (such that poste-

riors will always be proper). Nevertheless, our simulations in Section 4 indicate that our approach works well even in the case of quite weak instruments confirming simulations results of Conley et al. (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).

2.2 Nonparametric Effects

Since their introduction by Eilers and 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 et al. (2009); Reiss and Ogden (2009); Claeskens et al. (2009).

We will consider the Bayesian analogue to penalized splines as introduced by (Lang and Brezger, 2004). Thus, 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

and 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

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

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. We choose a conjugate inverse-gamma distribution with shape and scale parameters $a_{\tau_{r\ell}}$ and $b_{\tau_{r\ell}}$, i.e. $\tau^2_{r\ell} \sim \mathrm{IG}(a_{\tau_{r\ell}}, b_{\tau_{r\ell}})$.

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 $(\varepsilon_{1i}, \varepsilon_{2i}) \sim \mathrm{N}(0, \Sigma),\ i = 1, \ldots, n$ with constant covariance matrix Σ which is assumed to be a priori inverse-Wishart distributed $\Sigma \sim \mathrm{IW}(s_\Sigma, S_\Sigma)$ where IW denotes the inverted-Wishart distribution parameterized such that $\mathrm{E}(\Sigma) = S^{-1}_\Sigma/(s_\Sigma - 3)$. 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 $\mu_c,\ \Sigma_c,\ c = 1, \ldots, K^{**}$:

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

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. As a remedy, we consider a Gaussian Dirichlet Process Mixture (Escobar and 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 and 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

$$\theta_{k+1}|\theta_1, \ldots, \theta_k \sim \frac{\alpha}{\alpha + k} G_0 + \frac{1}{\alpha + k} \sum_{i=1}^{k} \delta_{\theta_i} \tag{5}$$

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 (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 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 et al. (2008) can be used (which is equally supported by our software). See Conley et al. (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 et al. (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 et al. (2008) propose to choose parameters $\mu_0,\ \tau_\Sigma,\ s_\Sigma$ 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 and 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 3.2. In order to express uncertainty about τ_Σ, Escobar and 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.

2.4 Hyperparameter Choices

From the properties of the inverse Wishart distribution (see e.g. Link and 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 et al. (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 et al. (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 $S_\Sigma = 0.2 I_2$ 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 et al. (2008) choose as default a $\mathrm{IG}(0.502, 0.085)$-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 paper.

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, $\tau_\Sigma \sim \mathrm{Ga}(a_\Sigma/2, b_\Sigma/2)$. We will follow Escobar and West (1995) and impose a diffuse gamma prior with default hyperparameters $a_\Sigma = 1$ and $b_\Sigma = 100$ which is in contrast to Conley et al. (2008) who choose a fixed τ_Σ. The impact of estimating τ_Σ versus fixing it is studied in our simulation study in Section 4.1.

With respect to the concentration parameter α, we follow the recommendation of Ishwaran and 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 $\tau_{r\ell}^2$ of nonparametric effects we choose the standard noninformative prior $\tau_{r\ell}^2 \sim \mathrm{IG}(0.001, 0.001)$ in the following.

3 Bayesian Inference

3.1 Estimation

Both equations (3) and (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 and 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 et al., 2011) of two Gibbs sampling algorithms with auxiliary variables given in Neal (2000). In addition, the implementation accompanying Conley et al. (2008) is integrated. Full details on all full conditionals are given in Appendix A.1.

3.2 Smoothing Parameter Estimation

In general, all nonparametric smoothing techniques involve some smoothing parameter controlling the roughness of the fit. This smoothing parameter has a strong impact on the estimate and has to be carefully chosen in finite samples. 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 $\mathrm{E}(\varepsilon_2|\varepsilon_1)$ and of the covariate functions have to be addressed differently. Here, smoothing parameter choice is particularly delicate 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 et al. (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 $\mathrm{E}(\varepsilon_2|\varepsilon_1)$. Marra and 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 4.2, 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 and 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 (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 are investigated in Section 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 and Greenberg (2010) and in Leslie et al. (2007). A rough diagnostic check is to visualise the estimated error density for discreteness.

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 and West, 1995) and thus the "smoothing parameter" of the conditional mean can be considered to be locally adaptive. See Escobar and West (1995) for connections between DPM and kernel density estimation with varying bandwidth.

3.3 Simultaneous Bayesian 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 simultaneous credible band as the Bayesian counterpart 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 the instrumental variable regression context Bayesian credible bands have the considerable advantage of naturally incorporating 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, we do not have to make any distributional assumption, i.e. also asymmetric bands can be obtained.

We follow Krivobokova et al. (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.

4 Simulations

4.1 Parametric Model

In this section, settings with linear covariate effects are simulated in order to compare the Bayesian approach to the well-established 2SLS estimator showing that it is capable of correcting endogeneity bias and that 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 et al. (2008) where normal and log-normal error distributions were simulated. We consider the basic model

$$y_2 = y_1 + z_2 + \varepsilon_2, \qquad y_1 = z_1 + z_2 + \varepsilon_1$$

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.

(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 and Gelman, 1991). Note that the degree of endogeneity varies over observations.

Densities of example draws from distributions in (iii) and (iv) are shown in Figure 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 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 et al. (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 $S_\Sigma = 0.2I_2$. 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.

In simulation settings (i) (Table 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 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 $R^2_{pop} = \frac{\mathrm{Var}(z_1)}{\mathrm{Var}(z_1)+\mathrm{Var}(z_2)+\sigma_1^2} = \frac{1/12}{1/12+1/12+\sigma_1^2} \approx 0.07$ and even slightly lower in setting (ii).

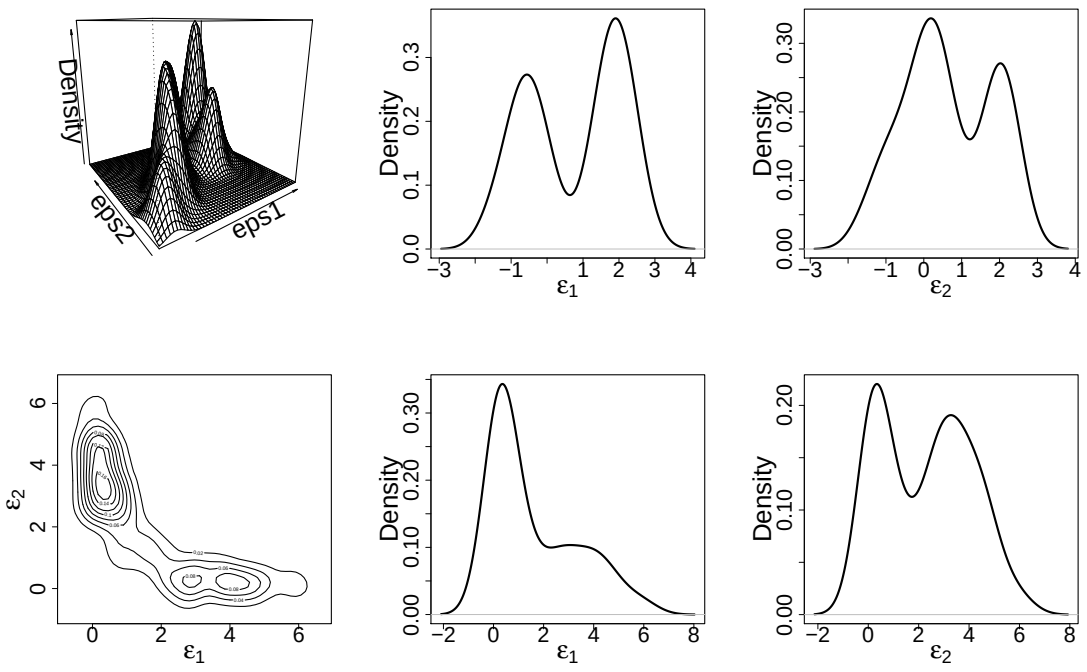

Figure 1: *Joint and marginal densities in one Monte Carlo draw of simulation setting (iii) (top panels) and setting (iv) (bottom panels).*

In settings (iii) and (iv) (Tables 3 and 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, 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.

Table 1: 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

Table 2: 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

4.2 Nonparametric Model

In our first two settings with nonparametric covariate effects, we replicate DGPs 1 and 4 of Su and Ullah (2008) aiming at getting some insight into the comparison of our Bayesian approach with Pinkse (2000)'s, Newey and Powell (2003)'s and Su and Ullah (2008)'s approaches. Moreover, we compare our results with Marra and Radice (2011)'s approach (extending the control function approach of Newey et al. (1999) to penalized splines). Thus, we consider settings

Table 3: 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

Table 4: 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

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

$$y_2 = \log(|y_1 - 1| + 1)\text{sgn}(y_1 - 1) + \varepsilon_2, \qquad y_1 = z_1 + \varepsilon_1$$

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

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

$$y_2 = 2\Phi(y_1) + \varepsilon_2, \qquad y_1 = \log(0.1 + z_1^2) + \varepsilon_1$$

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 $z_{1i} = 1 + 0.5z_{1,i-1} + 0.5v_z$.

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 and 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.

In Table 5, mean RMSEs and coverage rates of 95% simultaneous credible bands (when available) for DGP 1 and 4 of Su and 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 and 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 and 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 and 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 and Ullah (2008) (personal communication) chose the bandwidth h according to $h = c\sqrt{\text{Var}(y_1)}n^{-1/5}$

Table 5: Setting (a) and (b): DGPs of Su and 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 (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 (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 (Bayes)	0.122	0.976	0.069	0.984	0.061	0.992	0.037	0.992

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 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 approach is clearly outperformed by the DPM approach. Figure 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

Table 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 (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 (Bayes)	0.142	0.742	0.034	0.994	0.115	0.974

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 et al. (2010) in the single equation case. Note that for the control function approach as well as for the approaches compared in Su and Ullah (2008), no simultaneous confidence bands are available.

In summary, the proposed approach outperformed the control function approach based on GCV smoothing parameter selection and the estimators of Pinkse (2000), Newey and Powell (2003) and Su and Ullah (2008) (relying on the results given in Su and 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.

5 Class Size Effects on Student Achievements

In a very influential paper, Angrist and Lavy (1999) analyzed the effect of class size on 4th and 5th grades students tests scores in Israel. Among others they consider the model

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

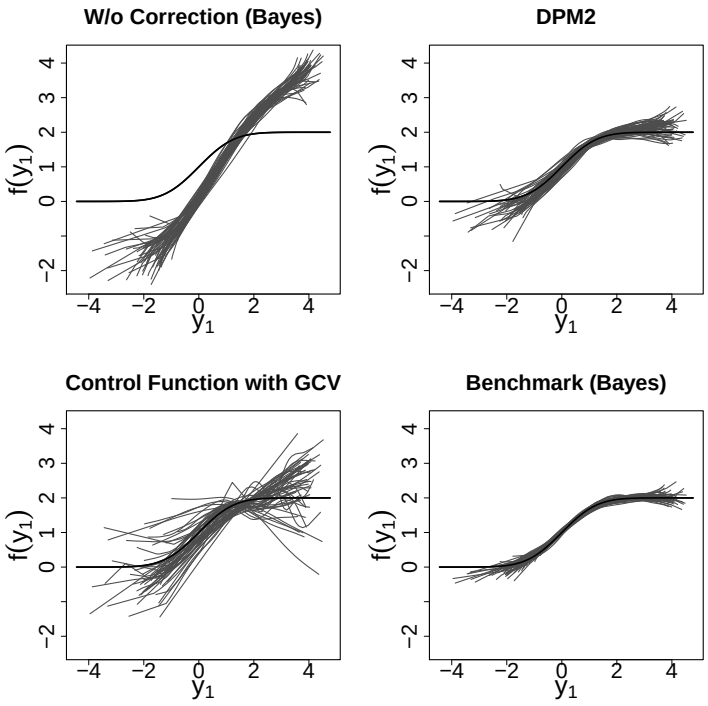

Figure 2: *Setting (b.iii): Estimated curves in first 50 simulation runs for* $n = 100$.

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.

To deal with the endogeneity of $csize_{ji}$ due to non-random assignment of class sizes, 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 int(k) is the largest integer less or equal to k.

Then, 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 and 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. This would correspond to a nonlinear effect, which could not properly be identified by a simple linear model. Thus, we relax the assumption of linear effects and extend the model of Angrist and Lavy (1999) to the following specification

$$tscore_{ji} = \gamma_{20} + f_{21}(csize_{ji}) + f_{22}(disadv_{ji}) + f_{23}(enrol_j) + \varepsilon_{2ji}, \quad (6)$$

$$csize_{ji} = \gamma_{10} + \gamma_{11} pcsize_{ji} + f_{12}(disadv_{ji}) + f_{13}(enrol_j) + \varepsilon_{1ji}. \quad (7)$$

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 3 shows estimated smooth effects for 4th graders (top panels) and 5th graders (bottom panels) in Equation (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.

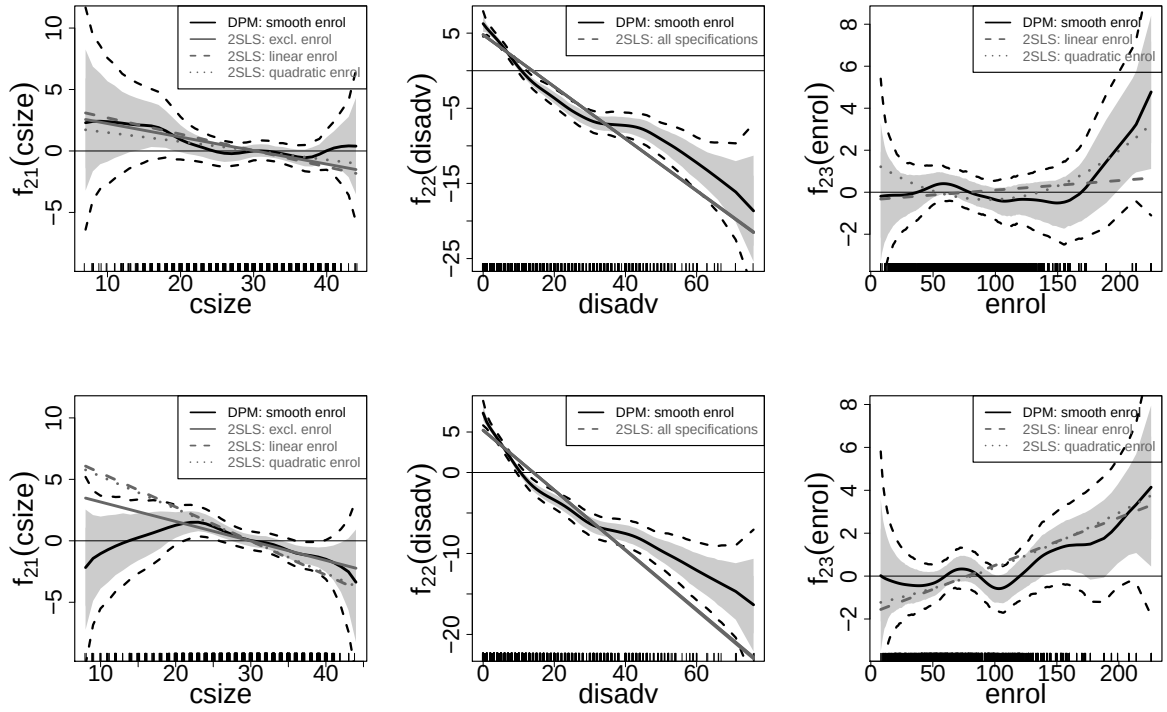

Figure 3: *Estimated effects for 4th (top) and 5th grade (bottom) students. Solid black lines show smooth curves in Equation (6) with 95% pointwise (gray areas) and simultaneous (areas between dashed lines) credible bands. 2SLS results for different parametric specifications of enrolment are given by gray lines.*

Regarding 4th grade students, no significant class size effect is found. This does not 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 than 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 conclusions on the impact of *csize* on students performance and its functional form. Note that

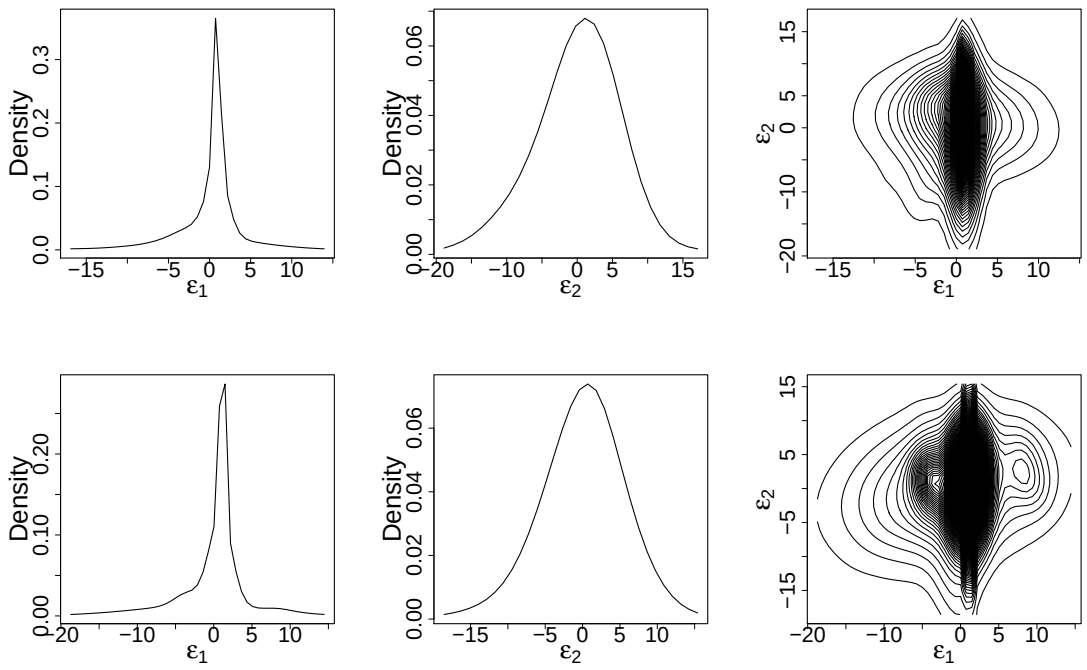

Figure 4: *Estimated marginal and joint error densities for 4th (top) and 5th grade (bottom) students.*

we find a significant deviation from the 2SLS fit (in models including enrolment). Also note that pointwise intervals (gray areas in Figure 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 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.

We find clearly nonnormal error densities in Figure 4. 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 and 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 $disadv = 1.5$).

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 and data-driven smoothing parameter selection which is particularly difficult in two-step frequentist approaches. 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. 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 et al., 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 and Kneib (2010). Our results can also be used for seemingly unrelated regression (SUR) extending Lang et al. (2003). The approach is implemented in the user-friendly R package *bayesIV*.

A Appendix

A.1 Full Conditionals

In the following, full conditionals for the parameters in the r-th equation, i.e. $r = 1$ for equation (4) and $r = 2$ for equation (3), are given.

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

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

with precision matrix

$$P_{\beta_{r\ell}} = X_{r\ell}^t \Sigma_{r|-r}^{-1} X_{r\ell} + \frac{\Delta_{r\ell}}{\tau_{r\ell}^2}$$

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

$$\mu_{\beta_{r\ell}} = P_{\beta_{r\ell}}^{-1} X_{r\ell}^t \Sigma_{r|-r}^{-1} (y_r - \widetilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r}))$$

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

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

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)

$$f_{r\ell}(\cdot) = (X_{r\ell}^t \Sigma_{r|-r}^{-1} X_{r\ell} + \frac{1}{\tau_{r\ell}^2}\Delta_{r\ell})^{-1} X_{r\ell}^t \Sigma_{r|-r}^{-1} (y_r - \widetilde{\eta}_r - \mathrm{E}(\varepsilon_r|\varepsilon_{-r})).$$

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 et al., 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

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

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

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

- 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

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

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 et al. (2008), α is drawn from a multinomial distribution. See Conley et al. (2008) for details.

References

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.

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 Dewatripont, M., Hansen, L., and Turnovsky, S., editors, *Advances in Economics and Econometrics: Theory and Applications, Eighth World Congress*, volume 2, pages 294–311. Cambridge University Press.

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

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 pe-

nalized spline estimators. *Biometrika*, 96(3):529–544.

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.

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

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*, 14:731–761.

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

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.

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.

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

Lang, S., Adebayo, S., Fahrmeir, L., and Steiner, W. (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.

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.

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.

Neal, R. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2):249–265.

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.

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.

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

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

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.

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