

ESTIMATION OF SEMIPARAMETRIC MODELS IN THE PRESENCE OF ENDOGENEITY AND SAMPLE SELECTION

Siddhartha Chib* Edward Greenberg Ivan Jeliazkov

June 2007

Abstract

We analyze a semiparametric model for data that suffer from the problems of incidental truncation, where some of the data are observed for only part of the sample with a probability that depends on a selection equation, and of endogeneity, where a covariate is correlated with the disturbance term. The introduction of nonparametric functions in the model permits significant flexibility in the way covariates affect response variables. We present an efficient Bayesian method for the analysis of such models that allows us to consider general systems of outcome variables and endogenous regressors that are continuous, binary, censored, or ordered. Estimation is computationally inexpensive as it does not require data augmentation for the missing outcomes, thus reducing computational demands and enhancing the mixing of the Markov chain Monte Carlo simulation algorithm. The methods are applied in a model of women's labor force participation and log-wage determination that accounts for endogeneity, incidental truncation, and non-linear covariate effects.

Keywords: Binary data; censored regression; data augmentation; incidental truncation; informative missingness; labor force participation; log-wage estimation; Markov chain Monte Carlo; model selection; sample selection; Tobit regression.

1 Introduction

In this article we discuss a model that contains three main components – sample selection, endogeneity, and nonparametric covariate effects. While each of these components by itself can lead to complications in estimation, their joint presence in a model brings out additional estimation challenges that require careful analysis.

The first of these model components, sample selection, is a common problem in applied studies. It arises when all variables are observed for a subset of the data—the *selected* sample—but only some of the variables are observed for the entire set of observational

**Address for correspondence:* Olin School of Business, One Brookings Drive, St. Louis MO 63130-4899; e-mail: chib@wustl.edu. Siddhartha Chib is Harry C. Hartkopf Professor of Econometrics and Statistics, Olin School of Business, Washington University in St. Louis; Edward Greenberg is Professor of Economics, Washington University in St. Louis; and Ivan Jeliazkov is Assistant Professor of Economics, University of California, Irvine.

units—the *potential* sample. In many cases, the factors that determine membership in the selected sample are correlated with those that determine the outcome. Such unobserved factors induce correlation between the outcomes and the selection mechanism and leads to “non-ignorable truncation” sometimes also called “incidental truncation” or “informative missingness.” In this case it is important to account for the mechanism that produces the selected sample from the potential sample. In early work, Heckman (1976, 1979) devised a well known and commonly applied two-step estimation procedure for a prototypical sample selection model. Many of the variants of the prototypical model are summarized in Wooldridge (2002) together with a number of alternatives to the two-step procedure; a comparison of several of these alternatives is given in Puhani (2000).

In addition to non-ignorable sample selection, the model we analyze includes endogenous covariates. A well-known example of endogeneity is the possibility that education, which affects wages, may be correlated with the disturbance in the wage equation because wages and education may both depend on such unobservable variables as ability or motivation. The standard approach for confronting endogeneity from both the frequentist and Bayesian viewpoints is via instrumental variables, which are variables that are uncorrelated with the error in the response variable and correlated with the endogenous covariate (for an extended summary, see Wooldridge, 2002, chapters 5 and 17).

We consider the third main regression component, nonparametric covariate effects, because the negative consequences of undetected nonlinearity can be significant. This can happen when the effect of an endogenous covariate is nonlinear but is mistakenly modeled linearly or as a low order polynomial. In this case, the conditional distribution of the responses, as well as the joint distribution of the errors, will appear non-Gaussian even if the true data generating process is Gaussian. For this reason, tests may erroneously reject the normality assumption when the real problem is undetected nonlinearity. Misspecified nonlinearity leads to misleading covariate effect estimates for all covariates because misspecification in one equation has ramifications for estimates of parameters, functions, and error distributions in other equations. These problems remain even if distribution-free estimation methods are used.

There is considerable Bayesian and frequentist work on relaxing the assumption that covariate effects are parametric. The underlying ideas can be traced as far back as Whit-

taker (1923) and are well summarized in Silverman (1985), Wahba (1990), and Hastie and Tibshirani (1990). Recent work that allows for nonparametric functions with endogenous variables, but without the complication of informative missingness, includes Chib and Greenberg (2006) and Hall and Horowitz (2005). Das, Newey, and Vella (2003) discuss series expansion (semi-nonparametric) estimators for a sample selection model with endogenous covariates, where the order of the expansion is allowed to increase as the sample size grows. They present two- and three-step estimators in the spirit of Heckman’s procedure. The paper does not develop asymptotic results for the case in which some of the endogenous covariates are incidentally truncated.

In the remainder of the paper we first present a hierarchical Bayesian model that accommodates the three components discussed above. A main goal is to allow the instrumental variables to have a nonparametric effect on the endogenous regressors which, in turn, may have a nonparametric effect on the response. Second, we present easily implementable simulation methods to fit the model. A point to note is that our estimation method proceeds without sampling for the missing outcomes. This allows for important computational advantages relative to an algorithm in which the missing data is included. We comment more on this below. Third, we address the problem of model choice by computing marginal likelihoods and Bayes factors to determine the posterior probabilities of competing models. Fourth, we report on the performance of our techniques in a simulation study and in the analysis of data on women’s labor supply. The data set includes log-wages for a sample of married women, which are missing for women who do not work. Education is modeled as an endogenous variable, and the selection mechanism is based on the hours worked, censored from below at zero. Section 7 offers concluding remarks.

2 The Model

Our model contains equations for a set of $J = J_1 + J_2$ variables, of which J_1 are observed only in the selected sample and J_2 other variables, including the selection variable, are always observed. Some of the variables may be endogenous covariates, and these may be observed only in the selected sample or for all units in the sample. To reduce notational complexity we describe the model and estimation methodology in detail for $J = 4$ response variables (y_1, \dots, y_4) , of which $J_1 = 2$ variables (y_1 and y_2) are observed only in the selected sample

and $J_2 = 2$ variables (y_3 and y_4) are always observed. The response variable y_1 may be viewed as the primary variable of interest. The variables y_2 and y_3 are endogenous regressors in the model for the primary response, and y_4 is the (censored) selection variable that is either zero or positive. The case of a binary selection variable is discussed in Section 3. The exogenous covariates, denoted by \mathbf{w} and \mathbf{x} , are assumed to be observed whenever the corresponding response variables they determine are observed. Special cases of the model that allow either or both of y_2 and y_3 to be absent do not require conceptual changes in the estimation algorithm, and generalizations to more variables are straightforward provided they do not lead to multicollinearity or the related problem of *concurvity* in nonparametric additive regression (Hastie and Tibshirani 1990).

In detail, for subject $i = 1, \dots, n$, the model we analyze is

$$y_{i1} = \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + g_1(y_{i2}, y_{i3}, \mathbf{w}_{i1}) + \varepsilon_{i1} \quad (1)$$

$$y_{i2} = \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + g_2(\mathbf{w}_{i2}) + \varepsilon_{i2} \quad (2)$$

$$y_{i3} = \mathbf{x}'_{i3}\boldsymbol{\beta}_3 + g_3(\mathbf{w}_{i3}) + \varepsilon_{i3} \quad (3)$$

$$y_{i4}^* = \mathbf{x}'_{i4}\boldsymbol{\beta}_4 + g_4(\mathbf{w}_{i4}) + \varepsilon_{i4}, \quad (4)$$

where the first equation models the primary response, the second and third equations are the marginal models for each of the endogenous regressors, and the fourth equation is the model for the latent censored selection variable y_{i4}^* (Tobin 1958); it is related to the observed selection variable by $y_{i4} = y_{i4}^* I(y_{i4}^* > 0)$, where $I(\cdot)$ is the indicator function. The vectors $\mathbf{x}_i = (\mathbf{x}_{i1}, \mathbf{x}_{i2}, \mathbf{x}_{i3}, \mathbf{x}_{i4})$ and $\mathbf{w}_i = (\mathbf{w}_{i1}, \mathbf{w}_{i2}, \mathbf{w}_{i3}, \mathbf{w}_{i4})$ are exogenous covariates, where the effects of \mathbf{x}_i are linear and those of \mathbf{w}_i are nonparametric. An important feature of this model is that the effects of the endogenous variables y_{i2}, y_{i3} on the primary response may be nonparametric. Correlation among $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})$ cause y_{i2} and y_{i3} , the covariates in the first equation, to be correlated with the error term ε_{i1} in that equation (endogeneity), and the presence of y_{i1} in the selected sample implies that y_{i4} is informative about ε_{i1} (“informative missingness” or “incidental truncation” or “sample selection”), implying that $E(\varepsilon_{i1} | y_{i2}, y_{i3}, y_{i4}, \mathbf{x}_{i1}, \mathbf{w}_{i1}) \neq 0$, so that endogeneity and sample selection must be formally taken into account in the modeling and estimation.

To model the unknown functions g_j in each of the $j = 1, \dots, J$ equations we assume the

additive nonparametric structure

$$g_j(\mathbf{s}_j) = \sum_{k=1}^{p_j} g_{jk}(s_{jk}), \quad (5)$$

which is discussed, for example, by Hastie and Tibshirani (1990), where p_j is the number of covariates in \mathbf{s}_j . The additive formulation is convenient because the “curse of dimensionality” renders nonparametric estimation of high-dimensional surfaces infeasible at present. Additive models are simple, easily interpretable, and sufficiently flexible for many practical applications. Additional flexibility can be attained by including covariate interactions in \mathbf{s}_j .

For identification reasons we assume that the covariates in $(\mathbf{x}_2, \mathbf{w}_2)$ contain at least one more variable than those included in $(\mathbf{x}_1, \mathbf{w}_1)$. These variables can be regarded as instrumental variables. Although identification in models with incidental truncation does not require instruments, we assume in our applications that $(\mathbf{x}_3, \mathbf{w}_3)$ contains covariates in addition to those in $(\mathbf{x}_1, \mathbf{w}_1)$. See Wooldridge (2002, chapters 5 and 17) for a discussion of the role of instrumental variables in estimating models with endogenous covariates and incidental truncation.

Several variants of this model can be specified: the selection variable can be binary (e.g., labor market status) rather than censored (e.g., hours of work), and the remaining endogenous variables may be censored, ordered, or binary. The endogenous variables y_2 and y_3 may enter the y_1 equation parametrically. We explain below how our methods can be applied to such cases. It is also straightforward to allow y_2 and y_3 to be vectors of endogenous variables.

The model is completed by assuming that the errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ are distributed as $\mathcal{N}_4(\mathbf{0}, \boldsymbol{\Omega})$, a four-variate normal distribution, where $\boldsymbol{\Omega}$ is an unrestricted symmetric positive definite matrix. It is possible to entertain other distributional forms for this joint distribution, but the normal assumption is an important case to consider because in many cases it provides the underpinning for more flexible distributions, such as finite mixture distributions or continuous scale mixture distributions.

A final point is that a normality assumption in conjunction with non-parametric functions is much more flexible than it may seem at first glance. Consider, for example, the case of a binary selection mechanism, where the marginal probit selection mechanism $P(y_{i4} = 1|g_4) = F(g_4(\mathbf{w}_{i4}))$ is fully flexible because $g_4(\cdot)$ is unrestricted, even though the

link function $F(\cdot)$ is the cdf of the Gaussian distribution. It is important to note that this way of modeling has advantages over modeling the distribution of the errors flexibly but considering only parametric effects in the mean of the selection equation; in this latter model, the effect of the \mathbf{x}_{i4} covariates is monotonic since F is monotonic and the mean is linear, but this is not necessarily the case when $g(\cdot)$ is nonparametric.

2.1 The Likelihood Function

We begin the development of our algorithm by discussing the likelihood function of the model in (1)–(4). Let $N_1 = \{i : s_i > 0\}$ be the n_1 observations in the selected sample, $N_2 = \{i : s_i = 0\}$ be the n_2 observations in the potential sample that are not in the selected sample, and $\boldsymbol{\theta}$ be the set of all model parameters and nonparametric functions. For the i th observation let

$$\mathbf{y}_{i1:2} = (y_{i1}, y_{i2})', \quad \mathbf{y}_{i1:3} = (y_{i1}, y_{i2}, y_{i3})', \quad \mathbf{y}_{i3:4} = (y_{i3}, y_{i4})', \quad \mathbf{y}_{i1:4} = (\mathbf{y}'_{i1:2}, \mathbf{y}'_{i3:4})',$$

and

$$\mathbf{y}_{i3:4}^* = (y_{i3}, y_{i4}^*)', \quad \mathbf{y}_{i1:4}^* = (\mathbf{y}'_{i1:2}, \mathbf{y}'_{i3:4}^*)',$$

where, as mentioned above, $y_{i4} = y_{i4}^* 1(y_{i4}^* > 0)$. The complete-data density function of the observations and latent data conditioned on $\boldsymbol{\theta}$ is given by

$$f(\mathbf{y}^*, \mathbf{y} | \boldsymbol{\theta}) = \prod_{i \in N_1} f(\mathbf{y}_{i1:4} | \boldsymbol{\theta}) \prod_{i \in N_2} f(\mathbf{y}_{i3:4}^* | \boldsymbol{\theta}) I(y_{i4}^* < 0) \quad (6)$$

because, for $i \in N_2$, only $\mathbf{y}_{i3:4}^*$ is observed and $\Pr(y_{i4} = 0 | y_{i4}^*, \boldsymbol{\theta}) = I(y_{i4}^* < 0)$. Now let

$$\mathbf{g}_{i1:2} = (g_1(y_{i2}, y_{i3}, \mathbf{w}_{i1}), g_2(\mathbf{w}_{i2}))', \quad \mathbf{g}_{i3:4} = (g_3(\mathbf{w}_{i3}), g_4(\mathbf{w}_{i4}))', \quad \mathbf{g}_{i1:4} = (\mathbf{g}'_{i1:2}, \mathbf{g}'_{i3:4})',$$

$$\mathbf{X}_{i3:4} = \begin{pmatrix} \mathbf{x}'_{i3} & 0 \\ 0 & \mathbf{x}'_{i4} \end{pmatrix}, \quad \mathbf{X}_{i1:4} = \begin{pmatrix} \mathbf{x}'_{i1} & 0 & 0 & 0 \\ 0 & \mathbf{x}'_{i2} & 0 & 0 \\ 0 & 0 & \mathbf{x}'_{i3} & 0 \\ 0 & 0 & 0 & \mathbf{x}'_{i4} \end{pmatrix},$$

and partition $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{pmatrix} \boldsymbol{\Omega}_{11} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \boldsymbol{\Omega}_{22} \end{pmatrix}.$$

Upon defining $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \boldsymbol{\beta}'_3, \boldsymbol{\beta}'_4)'$ and

$$\mathbf{J} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}, \quad (7)$$

such that $\mathbf{J}'\boldsymbol{\beta} = (\boldsymbol{\beta}'_3, \boldsymbol{\beta}'_4)'$, for $i \in N_1$ we have

$$f(\mathbf{y}_{i1:4}|\boldsymbol{\theta}) \propto |\boldsymbol{\Omega}|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y}_{i1:4} - \mathbf{g}_{i1:4} - \mathbf{X}_{i1:4}\boldsymbol{\beta})'\boldsymbol{\Omega}^{-1}(\mathbf{y}_{i1:4} - \mathbf{g}_{i1:4} - \mathbf{X}_{i1:4}\boldsymbol{\beta}) \right\},$$

and for $i \in N_2$

$$f(\mathbf{y}_{i3:4}^*|\boldsymbol{\theta}) \propto |\boldsymbol{\Omega}_{22}|^{-1/2} \exp \left\{ -\frac{1}{2}(\mathbf{y}_{i3:4}^* - \mathbf{g}_{i3:4} - \mathbf{X}_{i3:4}\mathbf{J}'\boldsymbol{\beta})'\boldsymbol{\Omega}_{22}^{-1}(\mathbf{y}_{i3:4}^* - \mathbf{g}_{i3:4} - \mathbf{X}_{i3:4}\mathbf{J}'\boldsymbol{\beta}) \right\}.$$

For some computations it is convenient to write the complete-data likelihood function of the observations and latent data as

$$f(\mathbf{y}^*, \mathbf{y}|\boldsymbol{\theta}) = \prod_{i \in N_1} f(\mathbf{y}_{i2:3}|\boldsymbol{\theta})f(y_{i1}|\mathbf{y}_{i2:3}, \boldsymbol{\theta})f(y_{i4}|\mathbf{y}_{i1:3}, \boldsymbol{\theta}) \prod_{i \in N_2} f(y_{i3}|\boldsymbol{\theta})f(y_{i4}^*|y_{i3}, \boldsymbol{\theta})I(y_{i4}^* < 0), \quad (8)$$

where all of the above densities are Gaussian, even though jointly they are not. As is standard in Tobit models, the likelihood function $f(\mathbf{y}|\boldsymbol{\theta}) \equiv f(\mathbf{y}_1, \mathbf{y}_2, \mathbf{y}_3, \mathbf{y}_4|\boldsymbol{\theta})$ is obtained by integrating $f(\mathbf{y}^*, \mathbf{y}|\boldsymbol{\theta})$ over the latent data y_{i4}^* , $i \in N_2$.

2.2 Prior distributions

We complete the model by specifying the prior distributions for the parameters and the nonparametric functions. We assume $\boldsymbol{\beta} = (\boldsymbol{\beta}'_1, \boldsymbol{\beta}'_2, \boldsymbol{\beta}'_3, \boldsymbol{\beta}'_4)'$ has a joint normal distribution with mean $\boldsymbol{\beta}_0$ and variance \mathbf{B}_0 and (independently) that the covariance matrix $\boldsymbol{\Omega}$ has an inverted Wishart distribution with parameters ν and \mathbf{Q} ,

$$\pi(\boldsymbol{\beta}, \boldsymbol{\Omega}) = \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0)\mathcal{IW}(\boldsymbol{\Omega}|\nu, \mathbf{Q}),$$

where $\mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0)$ is the density of the multivariate normal distribution and $\mathcal{IW}(\boldsymbol{\Omega}|\nu, \mathbf{Q})$ is that of the inverse Wishart distribution.

We model each of the unknown functions through the class of Markov process smoothness priors. This prior is easy to interpret, can approximate unknown functions arbitrarily well (with a penalty for “rough” functions), and has been widely used; see, for example, Shiller (1973, 1984), Gersovitz and MacKinnon (1978), Besag et al. (1995), Fahrmeir and Tutz (1997, Chapter 8), Müller et al. (2001), Fahrmeir and Lang (2001), Chib and Greenberg (2006), and Chib and Jeliazkov (2006).

Since the functions are assumed to be a priori independent, it is sufficient to give the details for (say) the function $g(w)$, where $g(\cdot)$ is any one of the univariate nonparametric

functions in (5) and w is the corresponding covariate; we have omitted the equation and function indices to simplify the notation. Since there may be repeated values of w , we define the $p \times 1$ *design point vector* $\mathbf{v} = (v_1, \dots, v_p)'$, $p \leq n$, with entries equal to the p unique ordered values of the w , $v_1 < \dots < v_p$, and also define $g_t = g(v_t)$.

Unrestricted additive models are identified only up to a constant because the likelihood remains unchanged if g_{jk} and g_{jh} , $k \neq h$, in (5) are simultaneously redefined as $g_{jk}^*(\cdot) = g_{jk}(\cdot) + a$ and $g_{jh}^*(\cdot) = g_{jh}(\cdot) - a$ for some constant a , so that $g_{jk}(\cdot) + g_{jh}(\cdot) = g_{jk}^*(\cdot) + g_{jh}^*(\cdot)$. To achieve identification, the nonparametric functions must be restricted to remove any free constants. We follow the approach of Shively et al. (1999) by restricting the functions to equal zero at the first ordered observation (i.e., $g_1 = 0$), allowing the parametric part of the model to absorb the overall intercept.

For the prior distribution of the second state g_2 of the process, we let

$$g_2 | \tau^2 \sim \mathcal{N}(g_{20}, \tau^2 G_0), \quad (9)$$

where τ^2 is a smoothness parameter discussed below. Individual τ^2 s, g_{20} s, and G_0 s are specified for each function, but this dependence is suppressed to simplify the notation.

We model the remaining function evaluations g_t , $t = 3, \dots, p$, as resulting from the realization of a second-order Markov process. Upon defining $h_t = v_t - v_{t-1}$, the second-order Markov process prior for g_t , $t = 3, \dots, p$, is

$$g_t = \left(1 + \frac{h_t}{h_{t-1}}\right) g_{t-1} - \frac{h_t}{h_{t-1}} g_{t-2} + u_t, \quad u_t \sim \mathcal{N}(0, \tau^2 h_t), \quad (10)$$

where τ^2 acts as a smoothness parameter in the sense that small values produce smooth functions and large values allow the function to interpolate the data more closely. The weights h_t in (10) adjust the variance to account for possibly irregular spacing between consecutive points in the design vector. Our weighting scheme assumes that the variance grows linearly with the distance h_t , a property satisfied by random walks, but other choices are possible (see e.g., Shiller 1984, Besag et al. 1995, and Fahrmeir and Lang 2001). We include τ^2 in the sampler, one for each function, with the prior distributions taking the inverse gamma form

$$\tau^2 \sim \mathcal{IG}(\nu_0/2, \delta_0/2),$$

with possibly different values of ν_0 and δ_0 for each τ^2 .

The prior specified by (9) and (10) implies a proper posterior distribution and yields a computationally convenient joint prior distribution for the unrestricted function evaluations $\mathbf{g} = (g_2, \dots, g_p)'$. To see the last point, note that after defining

$$\mathbf{H} = \begin{pmatrix} -\left(1 + \frac{h_3}{h_2}\right) & 1 & & & \\ \frac{h_4}{h_3} & -\left(1 + \frac{h_4}{h_3}\right) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_p}{h_{p-1}} & -\left(1 + \frac{h_p}{h_{p-1}}\right) & 1 \end{pmatrix}$$

and $\mathbf{\Sigma} = \text{diag}(G_0, h_3, \dots, h_p)$, the joint distribution of $\mathbf{g}|\tau^2$ can be expressed as

$$\mathbf{g}|\tau^2 \sim \mathcal{N}(\mathbf{g}_0, \tau^2 \mathbf{K}^{-1}), \quad (11)$$

where $\mathbf{g}_0 = \mathbf{H}^{-1}\tilde{\mathbf{g}}_0$ with $\tilde{\mathbf{g}}_0 = (g_{20}, 0, \dots, 0)'$ and the *penalty matrix* $\mathbf{K} = \mathbf{H}'\mathbf{\Sigma}^{-1}\mathbf{H}$. It is important to note that \mathbf{K} is banded and that manipulations of banded matrices require $O(n)$ operations, rather than the usual $O(n^3)$ for inversions or $O(n^2)$ for vector multiplication. Bandedness yields important computational savings in implementing an MCMC estimation algorithm. Finally, \mathbf{H} produces second order differences when post-multiplied by a vector, and its inverse constructs second order sums. In particular, the vector $\mathbf{H}^{-1}\tilde{\mathbf{g}}_0$ can be constructed by iterating (10) in expectation, starting with $g_{10} = 0$ and g_{20} from (9).

3 Estimation

Under the structure of the model, the posterior distribution for $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \boldsymbol{\Omega}, \{\mathbf{g}_j\}_{j=1}^q, \{\tau_j\}_{j=1}^q)$ and $(y_{i4}^*, i \in N_2)$ for the semiparametric model of (1)–(4) is given by

$$\begin{aligned} \pi(\boldsymbol{\theta}, (y_{i4}^*, i \in N_2) | \mathbf{y}) &\propto \mathcal{N}(\boldsymbol{\beta} | \boldsymbol{\beta}_0, \mathbf{B}_0) \mathcal{IW}(\boldsymbol{\Omega} | \nu, \mathbf{Q}) \\ &\times \left\{ \prod_{j=1}^q \mathcal{N}(\mathbf{g}_j | \mathbf{g}_{j0}, \tau_j^2 \mathbf{K}_j^{-1}) \mathcal{IG}(\tau_j^2 | \nu_{j0}/2, \delta_{j0}/2) \right\} \left\{ \prod_{i \in N_1} f(\mathbf{y}_{i1:4} | \boldsymbol{\theta}) \right\} \left\{ \prod_{i \in N_2} f(\mathbf{y}_{i3:4}^* | \boldsymbol{\theta}) I(y_{i4}^* < 0) \right\}. \end{aligned}$$

This posterior distribution can be summarized by MCMC methods (see Chib 2001 for a detailed survey of these methods). Letting $\boldsymbol{\theta} \setminus \boldsymbol{\theta}_j$ represent the elements of $\boldsymbol{\theta}$ except $\boldsymbol{\theta}_j$, we may summarize the algorithm as follows:

Algorithm 1: MCMC estimation of nonparametric incidental truncation model

1. Sample β from the distribution $\beta|\mathbf{y}^*, \boldsymbol{\theta} \setminus \beta$.
2. Sample $\boldsymbol{\Omega}$ from the distribution $\boldsymbol{\Omega}|\mathbf{y}^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}$ in a one-block, three-step procedure.
3. For $j = 1, \dots, p$, sample \mathbf{g}_j from the distribution $\mathbf{g}_j|\mathbf{y}^*, \boldsymbol{\theta} \setminus \mathbf{g}_j$.
4. For $j = 1, \dots, p$, sample τ_j^2 from the distribution $\tau_j^2|\mathbf{g}_j$.
5. For $i \in N_2$, sample y_{i4}^* from the distribution $y_{i4}^*|\mathbf{y}, \boldsymbol{\theta}$.

It is important to note that we do not involve the missing $\{\mathbf{y}_{i1:2}\}$ for $i \in N_2$ in this algorithm. This may seem surprising, because having the augmented “full” potential sample would reduce the model to a nonparametric seemingly unrelated regression (SUR) model that could be processed along the lines of Chib and Greenberg (1995) or Smith and Kohn (2000). Specifically, if the latent missing data are denoted by $\{\mathbf{y}_{i1:2}^\dagger\}$, sampling would proceed recursively by drawing from $[\boldsymbol{\theta}|\mathbf{y}^*, \{\mathbf{y}_{i1:2}^\dagger\}]$, $[\mathbf{y}^*|\boldsymbol{\theta}, \{\mathbf{y}_{i1:2}^\dagger\}]$, and $[\{\mathbf{y}_{i1:2}^\dagger\}|\mathbf{y}^*, \boldsymbol{\theta}]$, which would be updated by a series of full-conditional draws. We note, however, that our proposed approach is computationally easier and has three major advantages over the approach in which the missing data are part of the sampling. First, it reduces computational loads because time-consuming simulation of the missing $\{\mathbf{y}_{i1:2}^\dagger\}$ is not needed. Second, it improves the mixing of the Markov chain as sampling is not conditional on $\{\mathbf{y}_{i1:2}^\dagger\}$ – see Liu (1994) and Liu, Wong, and Kong (1994), who have extensively studied the collapsed Gibbs sampler and provide theoretical and practical evidence on the improved simulation performance that can be expected. These two advantages can become quite pronounced when there is a high proportion of missing outcomes or when the number of parameters is relatively large. Third, augmenting the sampler with the missing $\{\mathbf{y}_{i1:2}^\dagger\}$ is not straightforward if some of the covariates are missing when the corresponding responses are missing, which would require models for the missing covariates that are not necessary in our approach. We now turn to the details of the sampler.

Sampling β Our strategy for sampling $\beta|\boldsymbol{\theta} \setminus \beta, \mathbf{y}^*$ is to cast the system in the form of a SUR model using the definition of \mathbf{J} from (7). For $i \in N_1$ we have

$$\mathbf{y}_{i1:4}^* = \mathbf{X}_{i1:4}\boldsymbol{\beta} + \boldsymbol{\varepsilon}_i.$$

For $i \in N_2$ only $\mathbf{y}_{i3:4}^*$ is observed, and we multiply \mathbf{X}_i by \mathbf{J}' to select the relevant covariates. These observations may be combined with the data from N_1 and the prior distribution to yield $\boldsymbol{\beta}|\boldsymbol{\theta}\backslash\boldsymbol{\beta}, \mathbf{y}^* \sim \mathcal{N}(\mathbf{b}, \mathbf{B})$, where

$$\mathbf{b} = \mathbf{B} \left(\sum_{i \in N_1} \mathbf{X}'_{i1:4} \boldsymbol{\Omega}^{-1} \mathbf{y}_{i1:4}^* + \sum_{i \in N_2} \mathbf{J} \mathbf{X}'_{i3:4} \boldsymbol{\Omega}_{22}^{-1} \mathbf{y}_{i3:4}^* + \mathbf{B}_0^{-1} \mathbf{b}_0 \right)$$

$$\mathbf{B} = \left(\sum_{i \in N_1} \mathbf{X}'_{i1:4} \boldsymbol{\Omega}^{-1} \mathbf{X}_{i1:4} + \sum_{i \in N_2} \mathbf{J} \mathbf{X}'_{i3:4} \boldsymbol{\Omega}_{22}^{-1} \mathbf{X}_{i3:4} \mathbf{J}' + \mathbf{B}_0^{-1} \right)^{-1}.$$

As we mentioned above, this step does not require augmentation for the missing $\{\mathbf{y}_{i1:2}\}$, which reduces computations and improves the mixing of the chain because the sampling is from the marginal distribution rather conditional distribution given $\{\mathbf{y}_{i1:2}\}$.

Sampling $\boldsymbol{\Omega}$ The conditional distribution $\boldsymbol{\Omega}|\boldsymbol{\theta}\backslash\boldsymbol{\Omega}, \mathbf{y}^*$ is not an inverse Wishart distribution, because of the different forms of the complete data likelihood in the two subsamples N_1 and N_2 . Nonetheless, upon letting

$$\boldsymbol{\Omega}_{22} = \boldsymbol{\Omega}_{22} \tag{12}$$

$$\boldsymbol{\Omega}_{11 \cdot 2} = \boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21} \tag{13}$$

$$\mathbf{B}_{21} = \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\Omega}_{21}, \tag{14}$$

it is possible to derive and sample the distributions $\boldsymbol{\Omega}_{22}|\mathbf{y}^*, \boldsymbol{\theta}\backslash\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{11 \cdot 2}|\mathbf{y}^*, \boldsymbol{\theta}\backslash\boldsymbol{\Omega}_{11 \cdot 2}$, and $\mathbf{B}_{21}|\mathbf{y}^*, \boldsymbol{\Omega}_{11 \cdot 2}$, from which $\boldsymbol{\Omega}$ can be recovered. To show the form of these three conditional distributions we let

$$\boldsymbol{\eta}_{i1:4} = \mathbf{y}_{i1:4} - \mathbf{g}_{i1:4} - \mathbf{X}_{i1:4} \boldsymbol{\beta}, \quad i \in N_1$$

$$\boldsymbol{\eta}_{i3:4}^* = \mathbf{y}_{i3:4}^* - \mathbf{g}_{i3:4} - \mathbf{X}_{i3:4} \mathbf{J}' \boldsymbol{\beta}, \quad i \in N_2.$$

In terms of $\boldsymbol{\eta}_{i1:4}$ and $\boldsymbol{\eta}_{i3:4}^*$, the complete data likelihood is

$$\prod_{i \in N_1} f(\boldsymbol{\eta}_{i1:4}|\boldsymbol{\theta}) \prod_{i \in N_2} f(\boldsymbol{\eta}_{i3:4}^*|\boldsymbol{\theta}) \propto |\boldsymbol{\Omega}|^{-n_1/2} \exp \left[-\frac{1}{2} \sum_{i \in N_1} \boldsymbol{\eta}'_{i1:4} \boldsymbol{\Omega}^{-1} \boldsymbol{\eta}_{i1:4} \right]$$

$$\times |\boldsymbol{\Omega}_{22}|^{-n_2/2} \exp \left[-\frac{1}{2} \sum_{i \in N_2} \boldsymbol{\eta}'_{i3:4} \boldsymbol{\Omega}_{22}^{-1} \boldsymbol{\eta}_{i3:4}^* \right].$$

Partitioning the inverse Wishart hyperparameter matrix \mathbf{Q} conformably with $\mathbf{\Omega}$ as

$$\mathbf{Q} = \begin{pmatrix} \mathbf{Q}_{11} & \mathbf{Q}_{12} \\ \mathbf{Q}_{21} & \mathbf{Q}_{22} \end{pmatrix},$$

we find the posterior conditional distribution of $\mathbf{\Omega}$ to be

$$\begin{aligned} \pi(\mathbf{\Omega}|\mathbf{y}^*, \theta \setminus \mathbf{\Omega}) &\propto |\mathbf{\Omega}|^{-(\nu+n_1+J_1+J_2+1)/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{\Omega}^{-1}\mathbf{R}) \right] \\ &\times |\mathbf{\Omega}_{22}|^{-n_2/2} \exp \left[-\frac{1}{2} \text{tr} \left(\mathbf{\Omega}_{22}^{-1} \sum_{i \in N_2} \boldsymbol{\eta}_{i3:4}^* \boldsymbol{\eta}_{i3:4}' \right) \right], \end{aligned}$$

where $\mathbf{R} = \mathbf{Q} + \sum_{i \in N_1} \boldsymbol{\eta}_{i1:4}' \boldsymbol{\eta}_{i1:4}$.

Now making the change of variables from $\mathbf{\Omega}$ to $(\mathbf{\Omega}_{22}, \mathbf{\Omega}_{11.2}, \mathbf{B}_{21})$, with Jacobian $|\mathbf{\Omega}_{22}|^{J_1}$, we obtain the posterior distribution

$$\begin{aligned} \pi(\mathbf{\Omega}_{22}, \mathbf{\Omega}_{11.2}, \mathbf{B}_{21}|\mathbf{y}^*, \theta \setminus \mathbf{\Omega}) &\propto |\mathbf{\Omega}_{11.2}|^{-(\nu+n_1+J_1+J_2+1)/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{\Omega}^{-1}\mathbf{R}) \right] \\ &\times |\mathbf{\Omega}_{22}|^{-(\nu+n-J_1+J_2+1)/2} \exp \left[-\frac{1}{2} \text{tr} \left(\mathbf{\Omega}_{22}^{-1} \sum_{i \in N_2} \boldsymbol{\eta}_{i3:4}^* \boldsymbol{\eta}_{i3:4}' \right) \right], \end{aligned}$$

where we use $|\mathbf{\Omega}| = |\mathbf{\Omega}_{11.2}| |\mathbf{\Omega}_{22}|$. By the partitioned inverse theorem

$$\mathbf{\Omega}^{-1} = \begin{pmatrix} \mathbf{\Omega}_{11.2}^{-1} & -\mathbf{\Omega}_{11.2}^{-1} \mathbf{B}_{21}' \\ -\mathbf{B}_{21} \mathbf{\Omega}_{11.2}^{-1} & \mathbf{\Omega}_{22}^{-1} + \mathbf{B}_{21} \mathbf{\Omega}_{11.2}^{-1} \mathbf{B}_{21}' \end{pmatrix},$$

we are able to simplify the trace as

$$\begin{aligned} \text{tr}(\mathbf{\Omega}^{-1}\mathbf{R}) &= \text{tr}([\mathbf{\Omega}_{11.2}^{-1} \mathbf{R}_{11} - \mathbf{\Omega}_{11.2}^{-1} \mathbf{B}_{21}' \mathbf{R}_{21} - \mathbf{B}_{21} \mathbf{\Omega}_{11.2}^{-1} \mathbf{R}_{12} \\ &\quad + (\mathbf{\Omega}_{22}^{-1} + \mathbf{B}_{21} \mathbf{\Omega}_{11.2}^{-1} \mathbf{B}_{21}') \mathbf{R}_{22}]) \\ &= \text{tr}(\mathbf{\Omega}_{11.2}^{-1} [\mathbf{R}_{11} + \mathbf{B}_{21}' \mathbf{R}_{22} \mathbf{B}_{21} - \mathbf{B}_{21}' \mathbf{R}_{21} - \mathbf{R}_{12} \mathbf{B}_{21}]) + \text{tr}(\mathbf{\Omega}_{22}^{-1} \mathbf{R}_{22}) \\ &= \text{tr}(\mathbf{\Omega}_{11.2}^{-1} [(\mathbf{R}_{11} - \mathbf{R}_{12} \mathbf{R}_{22}^{-1} \mathbf{R}_{21}) + (\mathbf{B}_{21} - \mathbf{R}_{22}^{-1} \mathbf{R}_{21})' \mathbf{R}_{22} (\mathbf{B}_{21} - \mathbf{R}_{22}^{-1} \mathbf{R}_{21})]) \\ &\quad + \text{tr}(\mathbf{\Omega}_{22}^{-1} \mathbf{R}_{22}), \end{aligned}$$

where \mathbf{R} has been partitioned to conform to \mathbf{Q} . It now follows that

$$\begin{aligned} \pi(\mathbf{\Omega}_{22}, \mathbf{\Omega}_{11.2}, \mathbf{B}_{21}|\mathbf{y}^*, \theta \setminus \mathbf{\Omega}) &\propto |\mathbf{\Omega}_{11.2}|^{-(\nu+n_1+J_1+1)/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{\Omega}_{11.2}^{-1} \mathbf{R}_{11.2}) \right] \\ &\times |\mathbf{\Omega}_{11.2}|^{-J_2/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{\Omega}_{11.2}^{-1} (\mathbf{B}_{21} - \mathbf{R}_{22}^{-1} \mathbf{R}_{21})' \mathbf{R}_{22} (\mathbf{B}_{21} - \mathbf{R}_{22}^{-1} \mathbf{R}_{21})) \right] \\ &\times |\mathbf{\Omega}_{22}|^{-(\nu+n-J_1+J_2+1)/2} \exp \left[-\frac{1}{2} \text{tr}(\mathbf{\Omega}_{22}^{-1} [\mathbf{R}_{22} + \sum_{N_2} \boldsymbol{\eta}_{i3:4} \boldsymbol{\eta}_{i3:4}']) \right], \end{aligned}$$

where $\mathbf{R}_{11.2} = \mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$. From here we are able to conclude that

$$\begin{aligned}\boldsymbol{\Omega}_{22}|\mathbf{y}^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega} &\sim \text{IW}^{J_2}(\nu - J_1 + n_1 + n_2, \mathbf{Q}_{22} + \sum_{N_1, N_2} \boldsymbol{\eta}_{i3:4}^* \boldsymbol{\eta}_{i3:4}^*) \\ \boldsymbol{\Omega}_{11.2}|\mathbf{y}^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega} &\sim \text{IW}^{J_1}(\nu + n_1, \mathbf{R}_{11.2}) \\ \mathbf{B}_{21}|\mathbf{y}^*, \boldsymbol{\Omega}_{11.2} &\sim \text{MN}^{J_2 \times J_1}(\mathbf{R}_{22}^{-1}\mathbf{R}_{21}, \boldsymbol{\Omega}_{11.2} \otimes \mathbf{R}_{22}^{-1}).\end{aligned}$$

The sampling of $\boldsymbol{\Omega}$ can thus proceed from the respective full conditional densities of $\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{11.2}$, and \mathbf{B}_{21} , after which $\boldsymbol{\Omega}$ is recovered from (12)–(14).

Sampling the nonparametric functions We sample the nonparametric functions one at a time, conditional on all remaining functions, parameters, and the latent data, by exploiting efficient $O(n)$ sampling algorithms that utilize banded matrix operations. Our system contains four equations, where the k th equation contains q_k nonparametric functions. To simplify the notation, let \mathbf{g}_j denote the m_j -vector of unrestricted function evaluations of the j th function ($j = 1, \dots, q$, where $q = \sum_{k=1}^J q_k$) and let \mathbf{P}_j be an incidence matrix with entries $\mathbf{P}_j(h, l) = 1$ if $w_h = v_l$ and 0 otherwise; i.e., \mathbf{P}_j determines the correspondence between the vector of observations \mathbf{w}_j and \mathbf{v}_j , the vector of unique and ordered elements of \mathbf{w}_j . The l th component of $\mathbf{P}_j \mathbf{g}_j$ is $g_j(w_l)$. Because the sampling of each function is conditional on all parameters and other functions, in sampling the j th function we can focus only on the equation containing that function. Moving all components in that equation except $g_j(w_j)$ to the left-hand side, we can write

$$\eta_{ij} = y_{ik_j}^* - \mathbf{x}'_{ik_j} \boldsymbol{\beta}_{k_j} - \sum_{h \neq j} g_{k_j h}(w_{k_j h}) - E(\varepsilon_{ik_j} | \varepsilon_{i \setminus k_j}), \quad (15)$$

where subscript k_j indicates that the j th nonparametric function is in equation k_j . We define the vector $\boldsymbol{\eta}_j$, which has η_{ij} as its i th row; it is of dimension n_1 if equation k_j is part of the selected sample system and of dimension n if equation k_j is part of the sample that is always observed. It now follows from standard calculations that

$$\mathbf{g}_j | \mathbf{y}^*, \boldsymbol{\theta} \setminus \mathbf{g}_j \sim N(\hat{\mathbf{g}}_j, \hat{\mathbf{G}}_j)$$

where

$$\begin{aligned}\hat{\mathbf{G}}_j &= \left(\tau_j^{-2} \mathbf{K}_j + \mathbf{P}'_j \mathbf{V}_j^{-1} \mathbf{P}_j \right)^{-1} \\ \hat{\mathbf{g}}_j &= \hat{\mathbf{G}}_j \left(\tau_j^{-2} \mathbf{K}_j \mathbf{g}_{0j} + \mathbf{P}'_j \mathbf{V}_j^{-1} \boldsymbol{\eta}_j \right),\end{aligned}$$

and \mathbf{V}_j is a diagonal matrix with entries equal to $\text{Var}(\varepsilon_{ik_j}|\varepsilon_{\setminus ik_j})$, which introduces heteroskedasticity into the sampling of the unknown functions.

Sampling τ^2 The smoothness parameter for the j th nonparametric function is sampled from

$$\tau_j^2|\boldsymbol{\theta}\setminus\tau_j^2 \sim \mathcal{IG}\left(\frac{\nu_{j0} + m_j - 1}{2}, \frac{\delta_{j0} + (\mathbf{g}_j - \mathbf{g}_{j0})' \mathbf{K}_j (\mathbf{g}_j - \mathbf{g}_{j0})}{2}\right).$$

Sampling y_{i4}^* Following Chib (1992), this full conditional density is seen to be truncated normal:

$$y_{i4}^*|\mathbf{y}, \boldsymbol{\theta} \sim \mathcal{TN}_{(-\infty, 0)}(\mathbf{x}'_{i4}\boldsymbol{\beta}_4 + \mathbf{g}_4(\mathbf{w}_4) + E(\varepsilon_{i4}|\varepsilon_{i\setminus 4}), \text{Var}(\varepsilon_{i4}|\varepsilon_{i\setminus 4})), \quad i \in N_2.$$

3.1 Modifications for binary and censored variables and for longitudinal data

If there is more than one qualitative variable in the model—e.g., the response variable or one or more of the endogenous variables may be binary—three modifications to the basic scheme are required, as described in Albert and Chib (1993). First, y_{ik}^* is substituted for y_{ik} in the specification of the likelihood function, analogous to the use of y_{i4}^* in the selection equation. Second, y_{ik}^* is added to the sampler and sampled from an appropriately truncated distribution. Third, the variances of any binary or ordinal variables are set to one. If the response variable or any of the endogenous variables are censored, they are treated like y_{i4} in the discussion above with no variance restrictions.

The presence of binary or qualitative variables requires a modification of the algorithm to reflect the unit-variance constraints. When only one variable is binary or ordinal, the method for sampling $\boldsymbol{\Omega}$ may be utilized repeatedly to permit Gibbs sampling from inverse Wishart, Gaussian, and gamma distributions. A sampler for $\boldsymbol{\Omega}$ in which one of the variances is restricted to unity is presented in Munkin and Trivedi (2003) in a setting with endogeneity but without incidental truncation. When more than one variable is qualitative, the multiple unit-variance and positive definiteness restrictions on $\boldsymbol{\Omega}$ require a Metropolis-Hastings algorithm as in Chib and Greenberg (1998).

Finally, the estimation algorithm presented here can be extended to longitudinal settings by combining our algorithm with that of Chib and Carlin (1999) for continuous data settings or that of Chib and Jeliazkov (2006) for panels of discrete data.

4 Model Comparison

Bayesian model comparison based on posterior odds ratios or Bayes factors requires computation of the marginal likelihood for each model under consideration. The marginal likelihood of our model is given by the integral

$$m(\mathbf{y}) = \int f(\mathbf{y}|\boldsymbol{\beta}, \boldsymbol{\Omega}, \{\mathbf{g}_j\}, \{\tau_j^2\})\pi(\boldsymbol{\beta}, \boldsymbol{\Omega}, \{\mathbf{g}_j\}, \{\tau_j^2\}) d\boldsymbol{\beta} d\boldsymbol{\Omega} d\{\mathbf{g}_j\} d\{\tau_j^2\}.$$

We compute the marginal likelihood by the approach of Chib (1995), who points out that the multivariate integral can be estimated from the identity

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_j^*\}, \{\tau_j^{2*}\})\pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_j^*\}, \{\tau_j^{2*}\})}{\pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_j^*\}, \{\tau_j^{2*}\}|\mathbf{y})}, \quad (16)$$

where $\boldsymbol{\beta}^*$, $\boldsymbol{\Omega}^*$, $\{\mathbf{g}_j^*\}$, and $\{\tau_j^{2*}\}$ are fixed at high-density values. The posterior ordinate in the denominator of (16) must be estimated, but the likelihood and the prior ordinates in the numerator of (16) are directly available. To estimate the denominator of (16) we use the decomposition

$$\begin{aligned} \pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_j^*\}, \{\tau_j^{2*}\}|\mathbf{y}) = \\ \pi(\boldsymbol{\Omega}^*, \{\tau_j^{2*}\}|\mathbf{y}) \pi(\boldsymbol{\beta}^*|\mathbf{y}, \boldsymbol{\Omega}^*, \{\tau_j^{2*}\}) \prod_{i=1}^p \pi(\mathbf{g}_i^*|\mathbf{y}, \boldsymbol{\Omega}^*, \boldsymbol{\beta}^*, \{\tau_j^{2*}\}, \{\mathbf{g}_j^*\}_{j<i}), \end{aligned}$$

where the terms in the product are estimated by Rao-Blackwellization by averaging the full conditional densities defining the Gibbs sampler in Section 3 with respect to MCMC draws coming from appropriately structured MCMC runs. In particular, the first ordinate is estimated with draws from the main MCMC run, whereas the remaining ordinates in the product are evaluated with MCMC output from suitably defined reduced runs in which the parameters whose ordinates have already been obtained are held fixed and sampling is over the remaining elements of $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_J)$ and the latent data $\{y_{i4}^*\}$, i.e.:

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \{\boldsymbol{\theta}_j^*\}_{j<i}) = \frac{1}{G} \sum_{g=1}^G \pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \mathbf{y}_4^*, \{\boldsymbol{\theta}_j^*\}_{j<i}, \{\boldsymbol{\theta}_j^{(g)}\}_{j>i}),$$

for draws $\boldsymbol{\theta}_{j>i}^{(g)} \sim \pi(\boldsymbol{\theta}_{j>i}|\mathbf{y}, \boldsymbol{\theta}_{j<i}^*)$, $g = 1, \dots, G$ (see Chib 1995).

For the estimation of the marginal likelihood for semiparametric models, we make the following remarks. First, the numerical standard error of the marginal likelihood estimate,

which indicates the variation that can be expected if the simulation were to be repeated, can be calculated by the method in Chib (1995). Second, the choice of a suitable posterior density decomposition is very important in this model because it determines the balance between computational and statistical efficiency. The large dimension of $\{\mathbf{g}_j\}$, which may exceed the sample size, may increase the variability in the Rao-Blackwellization step if the full-conditional densities for the nonparametric functions are averaged over a conditioning set that changes with every iteration. For this reason these large-dimensional blocks should be placed towards the end of the decomposition so that more blocks in the conditioning set remain fixed. This strategy leads to higher statistical efficiency and comes at a reasonable computational cost since the sampling of the unknown functions is $O(n)$. Finally, note that the ordinate $\pi\left(\boldsymbol{\Omega}^*, \left\{\tau_j^{2*}\right\} \mid \mathbf{y}\right)$ is estimated jointly, rather than in individual reduced runs, because the parameters $\boldsymbol{\Omega}$ and τ_j^2 are conditionally independent given $\boldsymbol{\beta}$ and $\{\mathbf{g}_j\}$. This observation can reduce the computations significantly.

5 Simulation Study

To study the effectiveness of the sampler in estimating nonparametric functions we simulate data from a 3-equation version of the model in (1)–(4), where each equation contains an intercept and two additive functions:

$$y_{i1} = \beta_1 + g_{11}(y_{i2}) + g_{12}(w_{i1}) + \varepsilon_{i1}, \quad (17)$$

$$y_{i2} = \beta_2 + g_{21}(w_{i21}) + g_{22}(w_{i22}) + \varepsilon_{i2}, \quad (18)$$

$$y_{i3}^* = \beta_3 + g_{31}(w_{i31}) + g_{32}(w_{i32}) + \varepsilon_{i3}. \quad (19)$$

The nonparametric functions, graphed in Figure 1, have previously appeared in the literature: $g_{11}(v) = 2\Phi(v) - 1$, where $\Phi(\cdot)$ is the standard normal cdf; $g_{12}(v) = -0.8 + v + \exp(-30(v - 0.5)^2)$ for $v \in [0, 1]$; $g_{21}(v) = 1.5(\sin(\pi v))^2$ for $v \in [0, 1]$; $g_{22}(v) = \sin(v) + 1.5 \exp(-10v^2)$ for $v \in [-2, 4]$; $g_{31}(v) = 6(1 - \cos((\pi v/4)^2))$ for $v \in [0, 1]$; and $g_{32}(v) = 6v^3(1 - v^3)$ for $v \in [0, 1]$.

We assume that y_{i2} is observed for all i . Each function that depends on exogenous covariates is evaluated at $m = 51$ points chosen to be equally spaced on the support of each function, whereas g_{11} , which depends on the endogenous y_{i2} , is evaluated at points determined by the random realizations of y_{i2} . The data are generated with the equicorrelated

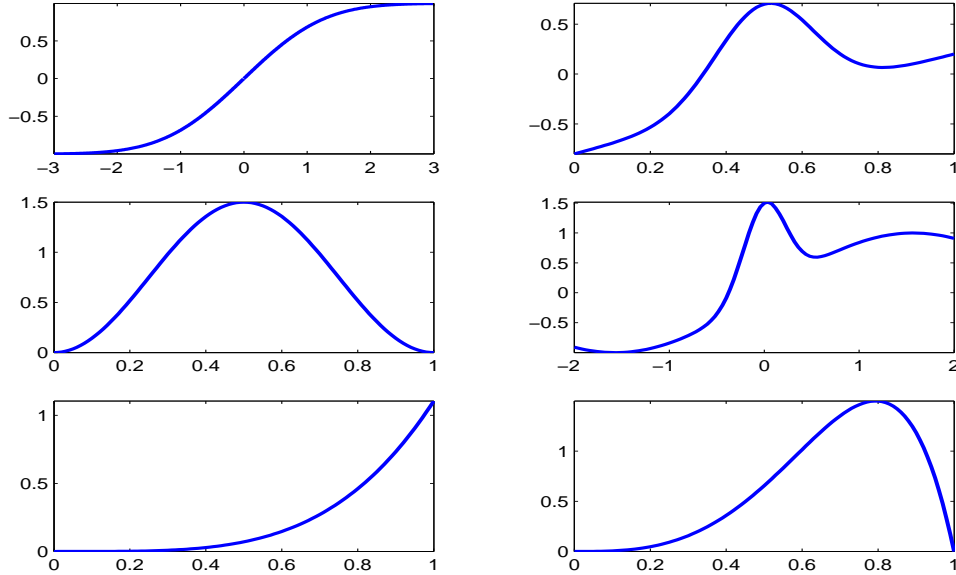


Figure 1: Functions used in the simulation study.

covariance matrix $\mathbf{\Omega} = (0.15\mathbf{I}_4 + 0.1\mathbf{ii}')$, which implies a relatively high degree of correlation (0.4) between the errors in the individual equations. Table 1 displays the signal-to-noise ratios for the functions, defined as the ratio of the range of the function to the standard deviation of the errors. These ratios vary from high noise (ratios around 2–3) to medium noise (ratios around 4–5). All else being equal (e.g., the functional form, the sample size, and the number of repeating observations per design point), the functions tend to be estimated more precisely as the signal-to-noise ratio is increased (cf. Wood, Kohn, Shively, and Jiang 2002). This study covers a reasonably “strong noise” scenario, and the performance of the techniques is likely to improve as the sample size is increased, more points are introduced per each design point, or the error variances are decreased.

| | Functions | | | | | |
|---|-----------|----------|----------|----------|----------|----------|
| | g_{11} | g_{12} | g_{21} | g_{22} | g_{31} | g_{32} |
| Range(g_{ij})/SD(ε_i) | 4.0 | 3.0 | 3.0 | 5.0 | 2.2 | 3.0 |

Table 1: Signal-to-noise ratios for the functions in the simulation study. (Note: the ratio for $g_{11}(\cdot)$ is an upper bound since the y_{i2} are randomly generated and subsequently censored, thus unlikely to fill the entire range.)

We summarize and report the performance of the sampler over 20 Monte Carlo replications for $n = 500, 1000,$ and 2000 . Due to randomness in obtaining the selected sample,

n_1 varies across the simulations. The ratio of the selected sample to the potential sample (n_1/n) varies from 0.816 to 0.866, with a median of 0.851 and a mean of 0.849. In all cases, we have used comparable priors: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 5 \times \mathbf{I})$, $\boldsymbol{\Omega} \sim \mathcal{IW}(J + 4, 1.2 \times \mathbf{I}_J)$, $g_{j2}|\tau_j^2 \sim \mathcal{N}\left(0, \tau_j^2/E\left(\tau_j^2\right)\right)$, for the six functions ($j = 1, \dots, 6$) in the study, $\tau_1^2 \sim \mathcal{IG}(6, .0004)$ and $\tau_j^2 \sim \mathcal{IG}(6, .04)$ for $j = 2, \dots, 6$; these priors imply that $E(\boldsymbol{\Omega}) = 0.4 \times \mathbf{I}$, $SD(\text{diag}(\boldsymbol{\Omega})) = 0.57 \times \mathbf{1}$, $E(\tau_1^2) = SD(\tau_1^2) = 0.0001$, $E(\tau_j^2) = SD(\tau_j^2) = 0.01$ for $j = 2, \dots, 6$, and $E_{\tau_j^2}(\text{Var}(g_{j2})) = 1$, for $j = 1, \dots, 6$. A tighter prior on the smoothness parameter was applied to the first function because all observations in \mathbf{y}_2 in the design were unique and the density is such that it assigns most of the observations around the middle of the cdf, which is approximately linear. As is well known in the literature, high values of the smoothing parameter τ^2 will lead to undersmoothing as the function becomes less smooth and tries to interpolate the observations, whereas low values of τ^2 lead to smoother functions. Of course, values of τ^2 that are too high or too low will yield poorer approximations that will tend to improve as more data become available. An example of the role of the smoothness parameter in over- and undersmoothing is presented in Chib and Jeliazkov (2006).

The posterior mean estimate $\hat{\mathbf{g}} = E\{g(\mathbf{v})|\mathbf{y}\}$, is found from MCMC runs of length 15000 following burn-ins of 2500 cycles. We gauge the performance of the method in fitting these functions by the root mean squared error, $\text{RMSE} = \sqrt{\frac{1}{m} \sum_{i=1}^m \{\hat{g}(v_i) - g(v_i)\}^2}$, where the true functions are shifted so as to satisfy the identification constraints of the model. Boxplots of the RMSEs for each function over the different samples are reported in Figure 2, where we see that the functions are estimated more precisely as the sample size grows. Function 1,

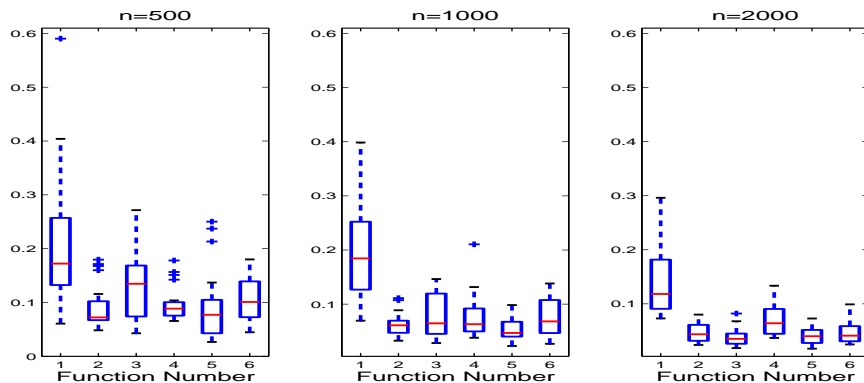


Figure 2: Boxplots of root mean squared errors of function estimates in the simulation study.

the nonparametric function of the endogenous covariate, is estimated less precisely than the other functions for two reasons: (1) all values in its design point vector are unique, whereas the other functions are evaluated at a smaller number of unique design point values, and (2) the first two functions are estimated from only n_1 observations, while the remaining functions are estimated from the full set of n observations. Finally, a close examination of the model structure shows that the level of $g_{11}(\cdot)$ in (17) is related not only to the intercept in its own equation, as is generally true for all functions in additive models (see the discussion on identification in Section 2.2), but is also correlated by construction with parameters in $\mathbf{\Omega}$. This can be seen by revisiting the discussion of the sampler for \mathbf{g} , especially equation (15), which shows that the errors in the endogenous covariate equation determine both the endogenous covariate and the conditional mean used in sampling g_{11} (through the covariance elements in $\mathbf{\Omega}$). For the other functions, the errors in other equations determine the conditional mean, but are not related to any other features of the sampled function such as its design point vector, which is what is special about $g_{11}(\cdot)$ in this setup. An example of the estimated functions for $n = 1000$ is given in Figure 3. Both Figures 2 and 3 show that the method recovers the true functions well.

We compute the inefficiency factors resulting from the Markov chain for the parametric components of the model. The inefficiency factor is defined as $1 + 2 \sum_{k=1}^L \rho_k(l)$, where $\rho_k(l)$ is the sample autocorrelation at lag l for the k th parameter in the sampling, with the summation truncated at values L at which the correlations taper off. This quantity may be interpreted as the ratio of the numerical variance of the posterior mean from the MCMC chain to the variance of the posterior mean from hypothetical independent draws. Figure 4, which displays the inefficiency factors obtained from the sampler discussed in Section 3, suggests several conclusions. First, although the inefficiency factors for β are the largest and do not seem to depend on the sample size, they are well within the limits found in the MCMC literature dealing with similar models (e.g. Chib and Greenberg 2006). For this reason, we emphasize that a longer MCMC chain may be required for more accurate estimation of β , but there are no other adverse consequences. Second, the elements of $\mathbf{\Omega}$ are sampled very efficiently (some are iid), and the parameters of $\mathbf{\Omega}$ that enter the Tobit equation (the last three elements of $\text{vech}(\mathbf{\Omega}) = (\omega_{11}, \omega_{21}, \omega_{22}, \omega_{31}, \omega_{32}, \omega_{33})$) are estimated better as sample sizes increase because there are more latent data points. Finally, since

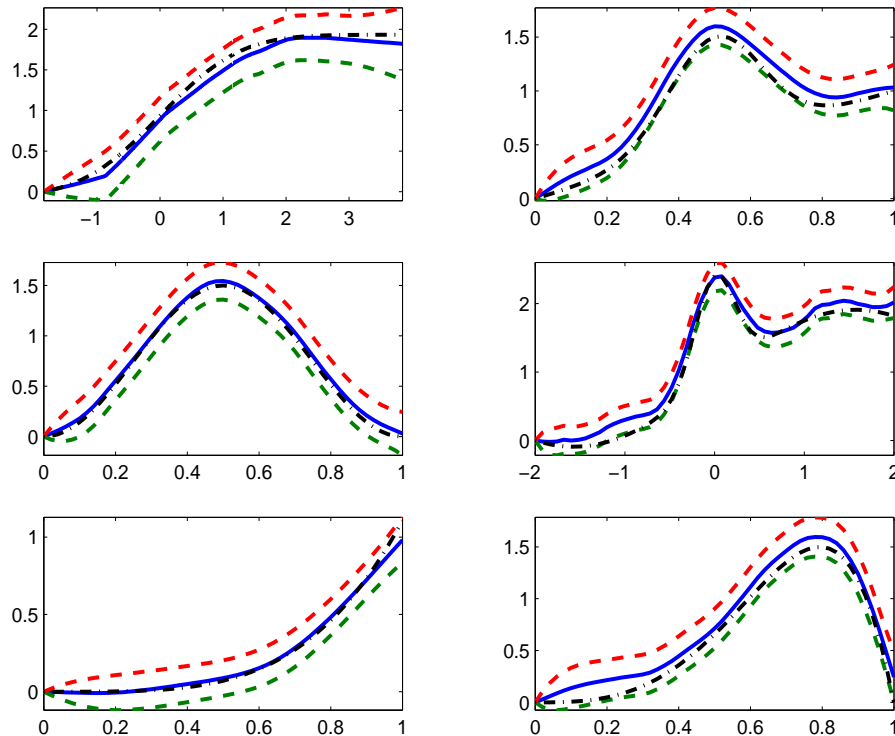


Figure 3: An example of function estimates in the simulation study: true functions (dot-dashes), estimated functions (solid lines), and pointwise confidence bands (dashes).

the estimates of the τ_j^2 depend on how well the corresponding functions are sampled, they depend on the sample size in a predictable way – as the sample size n grows and the functions $\{g_j\}$ are estimated better, so are the corresponding smoothness parameters $\{\tau_j\}$.

6 Estimating Women’s Wages

We apply the techniques of this paper to study the determinants of women’s wages, a topic that has been extensively studied because of the large increases in women’s participation in the labor force and in hours of work in the U. S. in the postwar period. Goldin (1989) reports a seven-fold increase in participation of married women since the 1920s, and Heckman (1993) underscores the importance of participation (entry and exit) decisions in estimating labor supply elasticities. The empirical analysis of women’s labor supply is complicated by the possible endogeneity of a covariate and by sample selection (incidental truncation) concerns. Endogeneity is likely to be a concern for the level of education—a covariate that

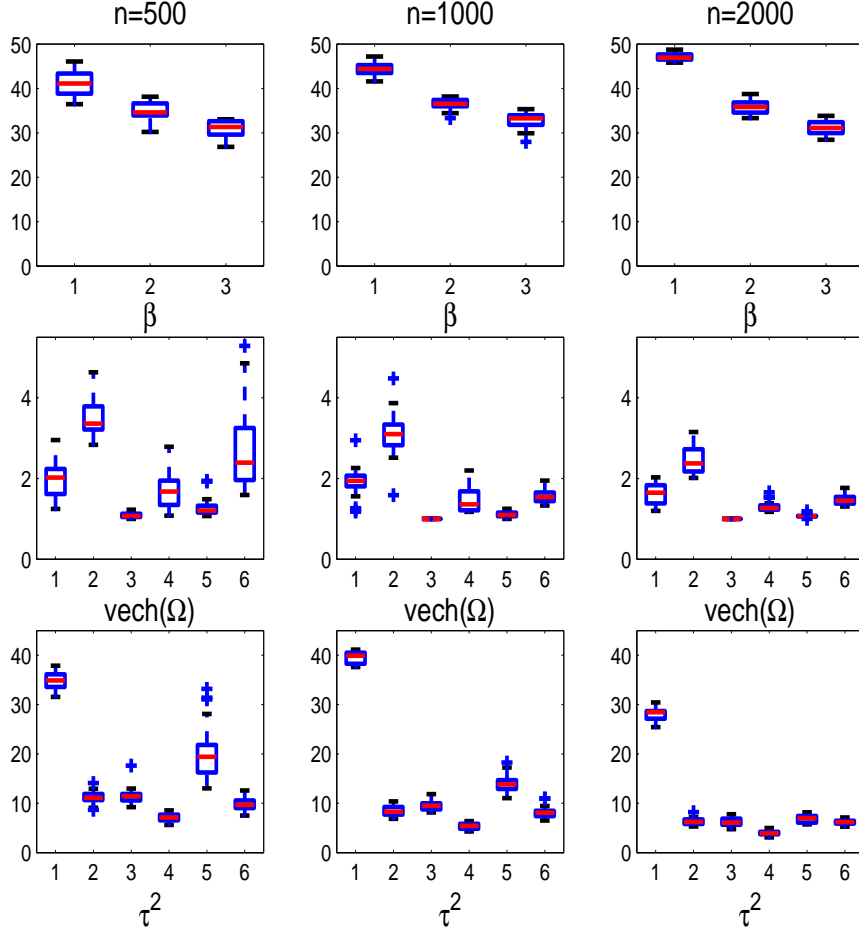


Figure 4: Inefficiency factors for the parameters of the model in the simulation study.

affects wages. Education is likely to be affected by some of the same unobserved variables (e.g., motivation, work ethic, perseverance, and intelligence) that are valued by employers in determining the wage rate. The problem of incidental truncation arises because wages are not observed for women who report zero annual hours of work, and such women may be out of the labor force because of excessively low wage offers. Our model allows for sample selection and endogeneity, and our nonparametric functions allow us to explore the presence of nonlinearities in the effects of some of the covariates. Such nonlinearities have been modelled by the inclusion of quadratic terms for certain variables (Mroz 1987 and Wooldridge 2002). We extend these results by estimating a semiparametric specification and comparing it to a number of alternatives models.

The data are from Mroz (1987). The potential sample consists of 753 married women,

| Variable | Explanation | Mean | SD |
|----------|--|--------|--------|
| WAGE | woman's wage rate (only for those working) | 4.18 | 3.31 |
| EDU | woman's years of schooling | 12.29 | 2.28 |
| HRS | woman's hours of work in 1975 | 740.58 | 871.31 |
| AGE | woman's age in years | 42.54 | 8.07 |
| EXPER | actual labor market experience in years | 10.63 | 8.07 |
| KLT6 | number of kids under 6 years old | 0.28 | 0.52 |
| KGE6 | number of kids 6–18 years old | 1.35 | 1.32 |
| NWINC | estimated nonwife income (1975, in \$10,000) | 2.01 | 1.16 |
| MEDU | mother's years of schooling | 9.25 | 3.37 |
| FEDU | father's years of schooling | 8.81 | 3.57 |
| HEDU | husband's years of schooling | 12.49 | 3.02 |

Table 2: Variables in the women's labor supply example from Mroz (1987). The sample consists of 753 married women, 428 of whom work. All summary statistics are for the full sample except where indicated.

428 of whom are employed and therefore in the selected sample. The variables in the data set are summarized in Table 2. We specify the econometric model as

$$y_{i1} = \mathbf{x}'_{i1}\boldsymbol{\beta}_1 + g_{11}(y_{i2}) + g_{12}(w_{i1}) + \varepsilon_{i1}, \quad (20)$$

$$y_{i2} = \mathbf{x}'_{i2}\boldsymbol{\beta}_2 + g_{21}(w_{i21}) + g_{22}(w_{i22}) + \varepsilon_{i2}, \quad (21)$$

$$y_{i3}^* = \mathbf{x}'_{i3}\boldsymbol{\beta}_3 + g_{31}(w_{i31}) + g_{32}(w_{i32}) + \varepsilon_{i3}, \quad (22)$$

where $y_{i3} = y_{i3}^*I(y_{i3}^* > 0)$ is the Tobit selection variable, so that y_{i1} is observed only when y_{i3} is positive, and y_{i2} is always observed. Based on the parametric models discussed in Mroz (1987) and Wooldridge (2002), we let

$$\begin{aligned} \mathbf{y}_i &= (y_{i1}, y_{i2}, y_{i3})' = (\ln(\text{WAGE}_i), \text{EDU}_i, \sqrt{\text{HRS}_i})', \quad \mathbf{x}_{i1} = 1 \\ \mathbf{x}_{i2} = \mathbf{x}_{i3} &= (1, \text{KLT6}_i, \text{KGE6}_i, \text{NWINC}_i, \text{MEDU}_i, \text{FEDU}_i, \text{HEDU}_i)' \\ \mathbf{w}_{i21} = \mathbf{w}_{i31} &= \text{AGE}_i, \quad \text{and} \quad \mathbf{w}_{i1} = \mathbf{w}_{i22} = \mathbf{w}_{i32} = \text{EXPER}_i. \end{aligned}$$

The choice of covariates and instruments in our setting does not deviate from earlier studies, but differs in the way in which the covariates are allowed to affect the responses. The nonparametric specification for AGE_i and EXPER_i is of interest because these covariates embody cohort, productivity, and life-cycle effects that are likely to affect wages nonlinearly. Wooldridge (2002, chapter 17) considers parametric models that contain linear and

quadratic terms in EXPER. The parameter estimates for our model are given in Table 3, and the nonparametric functions are plotted in Figure 5.

| Parameter | Covariate | Mean | SD | Median | Lower | Upper | Ineff |
|---------------|-----------|---------|--------|---------|---------|---------|--------|
| β_1 | 1 | 0.113 | 0.360 | 0.103 | -0.558 | 0.806 | 43.930 |
| β_2 | 1 | 4.817 | 0.395 | 4.821 | 4.037 | 5.597 | 12.388 |
| | KLT6 | 0.229 | 0.131 | 0.228 | -0.027 | 0.485 | 3.762 |
| | KGE6 | -0.084 | 0.056 | -0.084 | -0.193 | 0.024 | 2.172 |
| | NWINC | 0.144 | 0.059 | 0.145 | 0.030 | 0.259 | 1.699 |
| | MEDU | 0.134 | 0.023 | 0.134 | 0.090 | 0.178 | 1.000 |
| | FEDU | 0.093 | 0.021 | 0.094 | 0.051 | 0.135 | 1.000 |
| | HEDU | 0.347 | 0.023 | 0.347 | 0.301 | 0.393 | 1.421 |
| β_3 | 1 | -0.493 | 2.101 | -0.501 | -4.611 | 3.570 | 1.914 |
| | KLT6 | -9.636 | 1.586 | -9.642 | -12.727 | -6.512 | 2.084 |
| | KGE6 | -0.037 | 0.760 | -0.036 | -1.537 | 1.450 | 1.712 |
| | NWINC | -1.197 | 0.924 | -1.194 | -3.019 | 0.591 | 2.230 |
| | MEDU | 0.480 | 0.354 | 0.479 | -0.211 | 1.170 | 1.336 |
| | FEDU | 0.164 | 0.345 | 0.166 | -0.520 | 0.845 | 1.262 |
| | HEDU | -0.028 | 0.330 | -0.029 | -0.672 | 0.620 | 4.104 |
| ω_{11} | | 0.441 | 0.032 | 0.440 | 0.383 | 0.507 | 2.053 |
| ω_{21} | | 0.112 | 0.081 | 0.112 | -0.047 | 0.274 | 10.969 |
| ω_{22} | | 2.752 | 0.143 | 2.747 | 2.484 | 3.046 | 1.000 |
| ω_{31} | | -0.625 | 1.593 | -0.610 | -3.786 | 2.467 | 7.488 |
| ω_{32} | | 6.262 | 1.685 | 6.245 | 3.031 | 9.667 | 1.346 |
| ω_{33} | | 632.960 | 47.749 | 630.570 | 547.135 | 734.045 | 3.212 |
| τ_1^2 | | 0.008 | 0.005 | 0.006 | 0.003 | 0.020 | 4.176 |
| τ_2^2 | | 0.004 | 0.002 | 0.004 | 0.002 | 0.009 | 6.433 |
| τ_3^2 | | 0.005 | 0.003 | 0.004 | 0.002 | 0.012 | 6.083 |
| τ_4^2 | | 0.005 | 0.002 | 0.004 | 0.002 | 0.011 | 6.598 |
| τ_5^2 | | 0.011 | 0.009 | 0.008 | 0.003 | 0.035 | 11.550 |
| τ_6^2 | | 0.029 | 0.023 | 0.023 | 0.007 | 0.089 | 15.585 |

Table 3: Parameter estimates for nonparametric model of women’s wage function model under the priors $\beta \sim \mathcal{N}(\mathbf{0}, 5 \times \mathbf{I})$, $\Omega \sim \mathcal{IW}(7, 1.2 \times \mathbf{I})$, $g_{j2} | \tau_j^2 \sim \mathcal{N}\left(0, \tau_j^2 / E\left(\tau_j^2\right)\right)$ and $\tau_j^2 \sim \mathcal{IG}(6, .04)$ for $j = 1, \dots, 6$. The table also reports 95% confidence intervals and inefficiency factors from 25000 MCMC iterations.

The estimates in Table 3 are consistent with the predictions of economic theory. Results of the education equation reveal that the presence of younger children is associated with a higher level of mother’s education than having older children; presumably, having children earlier in life interfered with the woman’s education. The results also show that women who live in families with higher non-wife income, as well as women whose parents and

husband are better educated, are more likely to be better educated themselves. Results of the hours-worked equation suggest that having young children reduces the hours worked as evidenced by the negative mean and a 95% credibility interval that lies below zero, but older children have little impact on hours. Again, consistent with economic theory, higher non-wife income and lower parents' schooling reduce hours of work. The effect of husband's education is weak, both statistically and economically: its 95% credibility interval includes both negative and positive values, and its mean is small relative to its standard deviation.

The elements of Ω tell an interesting story. The estimates provide evidence that education is endogenous: the 95% credibility interval of ω_{21} is mostly in positive territory. In addition, even though the errors in the log wage equation are largely uncorrelated with those in the hours equation, sample selection is non-ignorable because the correlation between the errors in the education and hours equations is clearly positive.

We now consider the nonparametric functions plotted in Figure 5. The figure suggests that log-wages generally increase with education and experience, but that the increase is stronger for women with at least some college (the slope of g_{11} appears to change around 14 years of schooling). Moreover, the first 7–8 years of job experience lead to rapid gains in wages, after which wages appear to stabilize. An interesting nonlinearity appears at the end of the range of experience, where women with over 30 years of experience appear to command high wage rates. The amount of schooling does not vary with age for women between their 30s and 60s, when most people have completed school, but appears to be positively related to experience. Finally, the figure shows a strong negative effect of age on hours of work, which is consistent with the cohort and life-cycle effects, and a strong positive effect of experience on hours, which is consistent with increases in productivity as experience grows.

Since, with a few exceptions, the nonparametric profiles do not show substantial curvature, we compared this model to several simpler alternatives. The log-marginal likelihood of the model with six nonparametric functions is estimated to be -4236.44 with a numerical standard error of 0.144, and a model that includes only the first nonparametric function (the others being modelled linearly) had an almost identical log-marginal likelihood of -4236.43 with a numerical standard error of 0.075. These two semiparametric models therefore appear equiprobable, and the data do not provide enough information to distinguish between

them. These models were also compared to two parametric models—the log-marginal likelihood estimate for a linear model is -4237.405 with numerical standard error of 0.012 , and a parametric model that includes experience squared in all three equations has a log-marginal likelihood estimate of -4244.58 with numerical standard error of 0.012 . In this application it appears that linearity is a reasonable assumption for most of the covariate effects, but there is some evidence supporting the possibility that at least one and possibly two of the effects, namely those of (the endogenous covariate) education and experience, are nonlinear.

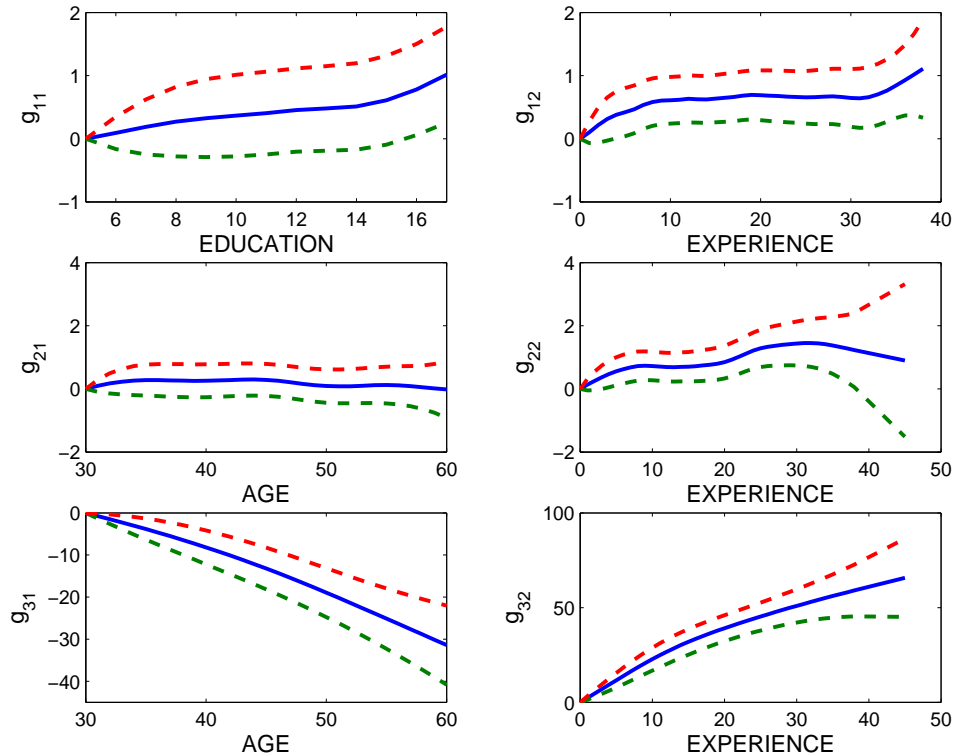


Figure 5: Function estimates in the log-wage application.

7 Conclusions

This paper has presented an efficient approach for analyzing a general class of models in which the problem of incidental truncation arises. The models include linear and nonparametric components and may involve endogenous regressors that enter the response equation nonparametrically. The class of models may involve multi-equation systems of responses

that comprise the selected sample or multi-equation systems that are always observed. The responses in these systems may be continuous, binary, ordered, or censored (Tobit).

An important aspect of the MCMC algorithm developed here for this class of models is that it does not require latent data for the responses that are unobserved. This feature of the estimation method enhances computational efficiency. Moreover, all sampling is from well-known full conditional distributions (Gibbs sampling) unless there are constraints on the covariance matrix arising from binary response or binary endogenous variables. In the latter cases, more general Metropolis-Hastings algorithms are available. The ability to compute marginal likelihoods makes it possible to compare different parametric and semiparametric model specifications in a fully Bayesian environment. A simulation study shows that the methods perform well, and an application involving a semiparametric model of women's labor force participation and log-wage determination illustrates that the model and the estimation methods are practical and can uncover interesting features in the data.

References

- ALBERT, J. and S. CHIB (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.
- BESAG, J., P. GREEN, D. HIGDON, and K. MENGENSEN (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.
- CHIB, S. (1992), "Bayes Inference in the Tobit Censored Regression Model," *Journal of Econometrics*, 51, 79–99.
- CHIB, S. (1995), "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.
- CHIB, S. (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Volume 5, eds J.J. Heckman and E. Leamer,) North Holland, Amsterdam, 3569–3649.
- CHIB, S. and B. CARLIN (1999), "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.
- CHIB, S. and E. GREENBERG (1995), "Hierarchical Analysis of SUR Models with Extensions to Correlated Serial Errors and Time Varying Parameter Models," *Journal of Econometrics*, 68, 339–360.
- CHIB, S. and E. GREENBERG (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361.

- CHIB, S. and E. GREENBERG (2006), “Analysis of Additive Instrumental Variable Models,” *Journal of Computational and Graphical Statistics*, in press.
- CHIB, S. and I. JELIAZKOV (2006), “Inference in Semiparametric Dynamic Models for Binary Longitudinal Data,” *Journal of the American Statistical Association*, 101, 685–700.
- DAS, M., W.K. NEWEY, and F. VELLA (2003), “Nonparametric Estimation of Sample Selection Models,” *Review of Economic Studies*, 70, 33–58.
- FAHRMEIR, L., and G. TUTZ (1997), *Multivariate Statistical Modelling Based on Generalized Linear Models*. New York: Springer-Verlag.
- FAHRMEIR, L., and S. LANG (2001), “Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors,” *Journal of the Royal Statistical Society, C*, 50, 201–220.
- GERSOVITZ, M. and J. MACKINNON (1978), “Seasonality in Regression: An Application of Smoothness Priors,” *Journal of the American Statistical Association*, 73, 264–273.
- GOLDIN, C. (1989), “Life-Cycle Labor-Force Participation of Married Women: Historical Evidence and Implications,” *Journal of Labor Economics*, 7, 20–47.
- HALL, P. and J.L. HOROWITZ (2005), “Nonparametric Methods for Inference in the Presence of Instrumental Variables,” *The Annals of Statistics*, 33, 2904–2929.
- HASTIE, T. and R. TIBSHIRANI (1990), *Generalized Additive Models*. New York: Chapman & Hall.
- HECKMAN, J.J. (1976), “The Common Structure of Statistical Models of Truncation, Sample Selection and Limited Dependent Variables,” *Annals of Economic and Social Measurement*, 15, 475–492.
- HECKMAN, J.J. (1979), “Sample Selection Bias as a Specification Error,” *Econometrica*, 47, 153–161.
- HECKMAN, J.J. (1993), “What Has Been Learned About Labor Supply in the Past Twenty Years?” *The American Economic Review*, 83, 116–122.
- LIU, J.S. (1994), “The Collapsed Gibbs Sampler in Bayesian Computations With Applications to a Gene Regulation Problem,” *Journal of the American Statistical Association*, 89, 958–966.
- LIU, J.S., W.H. WONG, and A. KONG (1994), “Covariance Structure of the Gibbs Sampler with Applications to the Comparisons of Estimators and Augmentation Schemes,” *Biometrika*, 81, 27–40.
- MROZ, T.A. (1987), “The Sensitivity of an Empirical Model of Married Women’s Hours of Work to Economic and Statistical Assumptions,” *Econometrica*, 55, 765–799.

- MÜLLER, P., G. ROSNER, L. INOUE, and M. DEWHIRST (2001): “A Bayesian Model for Detecting Acute Change in Nonlinear Profiles,” *Journal of the American Statistical Association*, 96, 1215–1222.
- MUNKIN, M.K., and P.K. TRIVEDI (2003), “Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: an application to the demand for healthcare,” *Journal of Econometrics*, 114, 197–220.
- PUHANI, P.A. (2000), “The Heckman Correction for Sample Selection and its Critique”, *Journal of Economic Surveys*, 14, 53–68.
- SHILLER, R. (1984): “Smoothness Priors and Nonlinear Regression,” *Journal of the American Statistical Association*, 79, 609–615.
- SHIVELY, T.S., R. KOHN, and S. WOOD (1999), “Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior” (with discussion), *Journal of the American Statistical Association*, 94, 777–806.
- SILVERMAN, B. (1985), “Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting” (with discussion), *Journal of the Royal Statistical Society, B*, 47, 1-52.
- SMITH, M. and R. KOHN (2000), “Nonparametric Seemingly Unrelated Regression,” *Journal of Econometrics*, 98, 257–281.
- TIERNEY, L. (1994), “Markov Chains for Exploring Posterior Distributions (with discussion),” *Annals of Statistics*, 22, 1701–1762.
- TOBIN, J. (1958), “Estimation of Relationships for Limited Dependent Variables,” *Econometrica*, 26, 24-36.
- WAHBA, G. (1978), “Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression,” *Journal of the Royal Statistical Society, B*, 40, 364-372.
- WHITTAKER, E. (1923), “On a New Method of Graduation,” *Proceedings of the Edinburgh Mathematical Society*, 41, 63-75.
- WOOLDRIDGE, J.M. (2002), *Econometric Analysis of Cross Section and Panel Data*. Cambridge: MIT Press.
- WOOD, S., and R. KOHN (1998), “A Bayesian Approach to Nonparametric Binary Regression”, *Journal of the American Statistical Association*, 93, 203–213.
- WOOD, S., KOHN, R., SHIVELY, T., and W. JIANG (2002), “Model Selection in Spline Nonparametric Regression,” *Journal of the Royal Statistical Society, Series B*, 64, 119–139.