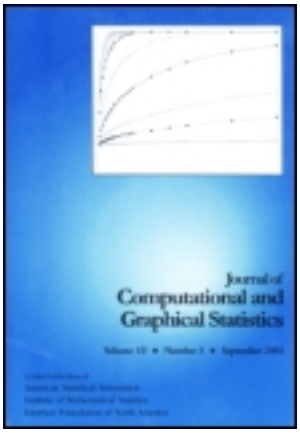

Journal of Computational and Graphical Statistics

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/ucgs20

Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection

Siddhartha Chib, Edward Greenberg and Ivan Jeliazkov

Siddhartha Chib is Harry C. Hartkopf Professor of Econometrics and Statistics, Olin Business School, Washington University in St. Louis, St. Louis, MO 63130 . Edward Greenberg is Professor Emeritus of Economics, Washington University in St. Louis, St. Louis, MO 63130 . Ivan Jeliazkov is Assistant Professor of Economics, University of California, Irvine, Irvine, CA 92697 .

Version of record first published: 01 Jan 2012.

To cite this article: Siddhartha Chib, Edward Greenberg and Ivan Jeliazkov (2009): Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection, Journal of Computational and Graphical Statistics, 18:2, 321-348

To link to this article: http://dx.doi.org/10.1198/jcgs.2009.07070

Supplemental materials for this article are available through the JCGS web page at http://www.amstat.org/publications.

Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection

Siddhartha CHIB, Edward GREENBERG, and Ivan JELIAZKOV

We analyze a semiparametric model for data that suffer from the problems of sample selection, 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 great 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 by Markov chain Monte Carlo (MCMC) methods. The algorithm we propose does not require simulation of the outcomes that are missing due to the selection mechanism, which reduces the computational load and improves the mixing of the MCMC chain. The approach is applied to a model of women's labor force participation and log-wage determination. Data and computer code used in this article are available online.

Key Words: Binary data; Censored regression; Data augmentation; Incidental truncation; Informative missingness; Labor force participation; Log-wage estimation; Markov chain Monte Carlo; Model selection; Tobit regression.

1. INTRODUCTION

In this article we extend the standard regression model by simultaneously allowing for sample selection, endogeneity, and nonparametric covariate effects. Although each of these issues by itself can lead to complications in estimation, their joint presence brings out additional estimation challenges that require careful analysis.

The problem of sample selection arises when all variables are observed for a subset of the observational units—the *selected* sample—but only some of the variables are observed for the entire set—the *potential* sample. The factors that determine membership in

Siddhartha Chib is Harry C. Hartkopf Professor of Econometrics and Statistics, Olin Business School, Washington University in St. Louis, St. Louis, MO 63130 (E-mail: *chib@wustl.edu*). Edward Greenberg is Professor Emeritus of Economics, Washington University in St. Louis, St. Louis, MO 63130 (E-mail: *edg@artsci.wustl.edu*). Ivan Jeliazkov is Assistant Professor of Economics, University of California, Irvine, Irvine, CA 92697 (E-mail: *ivan@uci.edu*).

Journal of Computational and Graphical Statistics, Volume 18, Number 2, Pages 321–348
DOI: 10.1198/jcgs.2009.07070

the selected sample are often correlated with those that determine the outcome. This sample selection problem is sometimes referred to as "nonignorable truncation," "incidental truncation," or "informative missingness."

We allow for endogenous covariates because such covariates are often the norm in observational studies. A well-known example of endogeneity occurs in the estimation of the effect of education on wages. In this case, it is possible that education may depend on such unobserved factors as ability and motivation that also have an effect on wages. Education is then an endogenous covariate and its effect on wages is confounded with the effect of ability and motivation on wages.

Finally, our model contains nonparametric covariate effects because the negative consequences of undetected nonlinearity on the estimation of covariate effects can be severe. If, for example, the effect of an endogenous covariate is mistakenly modeled linearly or as a low-order polynomial, the conditional distribution of the responses and the joint distribution of the errors may appear to be non-Gaussian even if the true data generating process is Gaussian. Misspecified nonlinearity can lead to misleading estimates for all covariate effects because misspecification in one equation can affect estimates of parameters, functions, and error distributions in other equations.

To illustrate a setting to which our model can be applied, consider data collected on a sample of married women who supply information about their hours of work, wage rates, on-the-job training, years of formal education, and such exogenous covariates as number and ages of children, age, and experience. This group is the potential sample. An important feature of these data is that information on wage rates and on-the-job training is available only for women reporting positive hours of work—these observations constitute the selected sample. The primary response variable of interest is log-wage rates, but on-the-job training and formal education are endogenous because they may be correlated with such unobservable determinants of log-wages as ability or motivation. Finally, there could be reason to believe that such covariates as age and years of education affect wages nonlinearly.

In early work, Heckman (1976, 1979) devised a two-step estimation procedure for a prototypical sample selection model. Many variants of the basic model were summarized in Wooldridge (2002) together with a number of alternatives to the two-step procedure; a comparison of several of these alternatives was given in Puhani (2000). The standard approach for confronting endogeneity from both the frequentist and Bayesian viewpoints is through the use of instrumental variables, which are variables that are uncorrelated with the error in the response variable and correlated with the endogenous covariate. For an extensive summary of these ideas, see Wooldridge (2002, chaps. 5 and 17).

There is considerable Bayesian and frequentist work on relaxing the assumption that covariate effects are parametric. The underlying ideas can be traced to Whittaker (1923) and are summarized in Wahba (1978), Silverman (1985), Hastie and Tibshirani (1990), Denison et al. (2002), and Wasserman (2006). Nonparametric techniques have been applied to multiequation systems with exogenous covariates such as the seemingly unrelated regression model in Smith and Kohn (2000), Holmes, Denison, and Mallick (2002), and Koop, Poirier, and Tobias (2005). These articles focused on estimation and model choice in a multiequa-

tion setting, which may include models for interaction between covariates, and we extend those techniques to account for sample selection issues and endogeneity. Recent work that allows for nonparametric functions with endogenous variables, but without the complication of informative missingness, includes Chib and Greenberg (2007) and Hall and Horowitz (2005). Gallant and Nychka (1987) discussed semi-nonparametric estimators for the basic sample selection model without endogenous covariates. The semi-nonparametric estimators are series expansions, where the order of the expansion is allowed to increase as the sample size grows. For a given sample size, this procedure is equivalent to approximating the covariate function by a polynomial. Similar series expansion estimators were considered in Das, Newey, and Vella (2003) for a model with endogenous covariates and sample selection. They presented two- and three-step estimators in the spirit of Heckman's procedure to study the impact of education on wages, but the model utilized a linear probability selection mechanism that does not constrain the probability of selection to be within [0, 1]. In addition, in finite samples the series expansion only spanned the class of low-order polynomials and the setup precluded the possibility of inference on how much smoothing is desirable because conventional optimal smoothing results do not apply. Moreover, they did not develop asymptotic results for the case in which some of the endogenous covariates are incidentally truncated. In contrast, we present a fully Bayesian, finite-sample inferential framework that accommodates endogeneity, sample selection, and flexible nonlinear effects within a formal probabilistic sample selection mechanism.

In Section 2, we present a hierarchical Bayesian model that accommodates the three components of the regression model discussed above. In Section 3, we present easily implemented simulation methods to fit the model, and, in Section 4, we address the problem of model choice by discussing the computation of marginal likelihoods and Bayes factors to determine the posterior probabilities of competing models. We report on the performance of our techniques in a simulation study in Section 5 and analyze an application dealing with the labor supply of married women in Section 6. 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, which 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 $J = 4$ variables $(y_1, \ldots, y_4)$, of which two (y_1 and y_2) are observed only in the selected sample and two (y_3 and y_4) are always observed. The response variable y_1 is 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 a Tobit (or censored) selection variable that is either zero or positive (see Tobin 1958). The case of a binary selection variable is discussed in Section 3.2. 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 to 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 $i = 1, \ldots, n$, the model we analyze is

$$y_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + g_1(y_{i2}, y_{i3}, \mathbf{w}_{i1}) + \varepsilon_{i1}, \tag{2.1}$$

$$y_{i2} = \mathbf{x}_{i2}'\boldsymbol{\beta}_2 + g_2(\mathbf{w}_{i2}) + \varepsilon_{i2}, \tag{2.2}$$

$$y_{i3} = \mathbf{x}_{i3}'\boldsymbol{\beta}_3 + g_3(\mathbf{w}_{i3}) + \varepsilon_{i3}, \tag{2.3}$$

$$y_{i4}^* = \mathbf{x}_{i4}'\boldsymbol{\beta}_4 + g_4(\mathbf{w}_{i4}) + \varepsilon_{i4}, \tag{2.4}$$

where the first equation models the primary response of interest y_{i1}, the second and third equations model the endogenous regressors y_{i2} and y_{i3}, and the fourth equation is the model for the latent censored selection variable y_{i4}^*. The latent y_{i4}^* is related to the observed selection variable y_{i4} by $y_{i4} = y_{i4}^* I(y_{i4}^* > 0)$, where $I(\cdot)$ is the indicator function. The selection variable y_{i4} determines the set of variables that are observed for the ith unit in the sample: if $y_{i4} > 0$, the entire vector $y_{i1:4} = (y_{i1}, y_{i2}, y_{i3}, y_{i4})'$ is observed, and when $y_{i4} = 0$, only $y_{i3:4} = (y_{i3}, y_{i4})'$ are observed, and $y_{i1:2} = (y_{i1}, y_{i2})'$ are missing. As an illustration, recall the married women's wage data example mentioned in the Introduction, where we let y_{i1} be the log-wage rate, y_{i2} be on-the-job training, y_{i3} be years of formal education, and y_{i4} be hours of work. 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 $\mathbf{x}_{ij}$ are linear and those of $\mathbf{w}_{ij}$ are nonparametric. An important feature of this model is that the effects of the endogenous variables y_{i2} and y_{i3} on the primary response y_{i1} may be nonparametric, although either or both can be entered linearly if desired. Correlation among $(\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3})$ causes y_{i2} and y_{i3} to be correlated with the error term ε_{i1} in that equation (endogeneity), and correlation between ε_{i1} and ε_{i4} implies that y_{i4} is informative about ε_{i1} (sample selection). As a result, endogeneity and sample selection must be formally taken into account in the modeling and estimation.

To model the unknown functions in the jth equation, we assume the additive nonparametric structure discussed, for example, by Hastie and Tibshirani (1990),

$$g_j(\mathbf{s}_j) = \sum_{k=1}^{q_j} g_{jk}(s_{jk}), \tag{2.5}$$

where s_{jk} is the kth covariate in $\mathbf{s}_j$ and q_j is the number of covariates in $\mathbf{s}_j$; the $g_{jk}(\cdot)$ are nonparametric functions described below. 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 the set of regressors modeled additively, as long as doing so does not lead to concurvity.

For identification reasons we assume that the covariates in $(\mathbf{x}_2, \mathbf{w}_2)$ and $(\mathbf{x}_3, \mathbf{w}_3)$ 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 $(\mathbf{x}_4, \mathbf{w}_4)$ contains covariates in

addition to those in $(\mathbf{x}_1, \mathbf{w}_1)$. 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. We explain later 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 $\varepsilon_i = (\varepsilon_{i1}, \varepsilon_{i2}, \varepsilon_{i3}, \varepsilon_{i4})'$ have a multivariate normal distribution $\mathcal{N}_4(\mathbf{0}, \mathbf{\Omega})$, where $\mathbf{\Omega}$ is an unrestricted symmetric positive definite matrix. It is possible to entertain other distributional forms for this joint distribution, but the normal assumption is important because 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 nonparametric functions is much more flexible than it may seem. For example, in the case of a binary selection mechanism, $\Pr(y_{i4} = 1|g_4) = F(g_4(\mathbf{w}_{i4}))$, the marginal probit selection mechanism is fully flexible because $g_4(\cdot)$ is unrestricted, even though the link function $F(\cdot)$ is the cdf of the Gaussian distribution. 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 the latter model, the effect of the $\mathbf{x}_4$ covariates is monotonic because F is monotonic and the mean is linear, but this is not necessarily the case when $g_4(\cdot)$ is nonparametric.

2.1 THE LIKELIHOOD FUNCTION

We begin the development of our algorithm with a discussion of the likelihood function of model (2.1)–(2.4). For the computations that follow, we define the vectors

$$\mathbf{y}^*_{i3:4} = (y_{i3}, y^*_{i4})', \qquad \mathbf{y}^*_{i1:4} = (y_{i1}, y_{i2}, y_{i3}, y^*_{i4}),$$
$$\mathbf{g}_{i1:2} = (g_1(y_{i2}, y_{i3}, \mathbf{w}_{i1}), g_2(\mathbf{w}_{i2}))', \qquad \mathbf{g}_{i3:4} = (g_3(\mathbf{w}_{i3}), g_4(\mathbf{w}_{i4}))',$$
$$\mathbf{g}_{i1:4} = (\mathbf{g}'_{i1:2}, \mathbf{g}'_{i3:4})',$$
$$\mathbf{X}_{i3:4} = \begin{pmatrix} \mathbf{x}'_{i3} & 0 \\ 0 & \mathbf{x}'_{i4} \end{pmatrix}, \qquad \mathbf{X}_{i1:4} = \begin{pmatrix} \mathbf{x}'_{i1} & 0 & 0 & 0 \\ 0 & \mathbf{x}'_{i2} & 0 & 0 \\ 0 & 0 & \mathbf{x}'_{i3} & 0 \\ 0 & 0 & 0 & \mathbf{x}'_{i4} \end{pmatrix}.$$

We also set $N_1 = \{i : y_{i4} > 0\}$ to be the n_1 observations in the selected sample and $N_2 = \{i : y_{i4} = 0\}$ to be the n_2 observations in the potential sample that are not in the selected sample. These definitions imply that N_1 is the set of indices for which we observe all four variables in $y_{i1:4}$, and N_2 is the set for which we observe only $y_{i3:4}$. Finally, let $\boldsymbol{\theta}$ be the set of all model parameters and nonparametric functions.

The complete-data density function of the observations and latent data conditioned on $\boldsymbol{\theta}$ is given by

$$f(\mathbf{y}, \mathbf{y}^*_4|\boldsymbol{\theta}) = \left[\prod_{i\in N_1} f(\mathbf{y}_{i1:4}|\boldsymbol{\theta})\right]\left[\prod_{i\in N_2} f(\mathbf{y}^*_{i3:4}|\boldsymbol{\theta}) I(y^*_{i4} < 0)\right], \tag{2.6}$$

where $\mathbf{y}$ contains $\mathbf{y}_{i1:4}$ for $i \in N_1$ and $\mathbf{y}_{i3:4}$ for $i \in N_2$, whereas $\mathbf{y}_4^*$ contains the observations on y_{i4}^* for $i \in N_2$. The second product on the right side of (2.6) is derived from

$$f(y_{i3}, y_{i4}, y_{i4}^*|\boldsymbol{\theta}) = f(y_{i3}, y_{i4}^*|\boldsymbol{\theta}) \Pr(y_{i4} = 0|y_{i3}, y_{i4}^*, \boldsymbol{\theta}),$$

because $y_{i4} = 0$ for $i \in N_2$, and $\Pr(y_{i4} = 0|y_{i3}, y_{i4}^*, \boldsymbol{\theta}) = I(y_{i4}^* < 0)$. Now partition $\boldsymbol{\Omega}$ as

$$\boldsymbol{\Omega} = \begin{pmatrix} \underset{(2\times 2)}{\boldsymbol{\Omega}_{11}} & \boldsymbol{\Omega}_{12} \\ \boldsymbol{\Omega}_{21} & \underset{(2\times 2)}{\boldsymbol{\Omega}_{22}} \end{pmatrix}; \tag{2.7}$$

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

$$\mathbf{J} = \begin{pmatrix} \mathbf{0} \\ \mathbf{I} \end{pmatrix}, \tag{2.8}$$

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

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

and for $i \in N_2$,

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

which gives us the terms needed in (2.6). The above decomposition of $f(\mathbf{y}, \mathbf{y}_4^*|\boldsymbol{\theta})$ is the basis for the straightforward and efficient sampler that we utilize in the remainder of the article, but some computations require the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ marginally of the latent $\mathbf{y}_4^*$. For those cases, it is convenient to write the complete-data likelihood function of the observations and latent data as

$$\begin{aligned} f(\mathbf{y}, \mathbf{y}_4^*|\boldsymbol{\theta}) = &\left[\prod_{i\in N_1} f(\mathbf{y}_{i2:3}|\boldsymbol{\theta}) f(y_{i1}|\mathbf{y}_{i2:3}, \boldsymbol{\theta}) f(y_{i4}|\mathbf{y}_{i1:3}, \boldsymbol{\theta})\right] \\ &\times \left[\prod_{i\in N_2} f(y_{i3}|\boldsymbol{\theta}) f(y_{i4}^*|y_{i3}, \boldsymbol{\theta}) I(y_{i4}^* < 0)\right], \end{aligned} \tag{2.9}$$

where all densities are Gaussian, even though jointly they are not. As is standard in censored-data models, the likelihood function $f(\mathbf{y}|\boldsymbol{\theta})$ is obtained by integrating $f(\mathbf{y}, \mathbf{y}_4^*|\boldsymbol{\theta})$ over the latent data $\mathbf{y}_4^*$, which is easily accomplished for the decomposition in (2.9) because for each $i \in N_2$ such integration only involves computing the cdf of a univariate Gaussian distribution.

2.2 PRIOR DISTRIBUTIONS

We complete the model by specifying the prior distributions for the parameters and the nonparametric functions. We assume that $\boldsymbol{\beta}$ 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 the density 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, chap. 8), Müller et al. (2001), Fahrmeir and Lang (2001), Chib and Jeliazkov (2006), and Chib and Greenberg (2007). We note that other nonparametric modeling approaches could be applied in our setting without altering our way of dealing with sample selection and endogeneity issues. These include the integrated Wiener process priors proposed in Wahba (1978) and applied within Bayesian inference by Wood and Kohn (1998) and Shively, Kohn, and Wood (1999). Other approaches include regression splines (Smith and Kohn 1996), B-spline priors (Silverman 1985), and wavelets (Denison et al. 2002). Recent extensions of these techniques to Bayesian free-knot spline modeling were given in Denison, Mallick, and Smith (1998) and DiMatteo, Genovese, and Kass (2001). Extensions to multivariate functions were presented in Wood et al. (2002) and Hansen and Kooperberg (2002). See Wasserman (2006) for a discussion of different nonparametric modeling approaches from the frequentist viewpoint or Denison et al. (2002) for a Bayesian perspective.

Although alternative nonparametric modeling approaches can be pursued for the unknown functions, two issues that can arise when alternatives are considered should be kept in mind. First, computations involving some nonparametric approaches, for example, the integrated Wiener process prior approach, may involve $O(n^3)$ operations, an important consideration for large samples. Although adaptive approaches to knot selection (see Denison, Mallick, and Smith 1998 and DiMatteo, Genovese, and Kass 2001) can reduce the computational burden, these approaches are also costly to implement because selecting the number and location of knots is nontrivial. Second, many classes of priors used in nonparametric functional modeling lead to partially improper priors on the unknown functions that are problematic for model choice. The approach we follow leads to an efficient estimation algorithm that requires $O(n)$ rather than $O(n^3)$ computations and produces proper priors for the unknown functions and parameters.

Because the unknown functions are treated similarly and assumed to be a priori independent, it is sufficient to give the details for any one of the univariate nonparametric functions in (2.5). To understand the main modeling issues in our setting, consider the jth equation, where interest focuses on the kth function $g_{jk}(\cdot)$ in that equation ($j = 1, \ldots, J$, $k = 1, \ldots, q_j$). If $j \leq J_1$ (i.e., equation j is observed only in the selected sample), the vector of covariates $\mathbf{w}_{jk}$ that enters $g_{jk}(\cdot)$ contains n_1 observations, that is, $\mathbf{w}_{jk} = (w_{1jk}, \ldots, w_{n_1 jk})'$ consisting of $\{w_{ijk}\}$ for $i \in N_1$; otherwise, when $J_1 < j \leq J$ (i.e., equation j is always observed), we have $\mathbf{w}_{jk} = (w_{1jk}, \ldots, w_{njk})'$.

Because there may be repeated values in $\mathbf{w}_{jk}$, we define the $p_{jk} \times 1$ *design point vector* $\mathbf{v}_{jk} = (v_{jk,1}, \ldots, v_{jk,p_{jk}})'$ of unique ordered values of $\mathbf{w}_{jk}$ with $v_{jk,1} < \cdots < v_{jk,p_{jk}}$, where $p_{jk} \leq n_1$ if $j \leq J_1$ or $p_{jk} \leq n$ if $J_1 < j \leq J$. The notation used for the elements in $\mathbf{v}_{jk}$ is intended to emphasize that enumeration in that vector does not correspond to the

enumeration in $\mathbf{w}_{jk}$, whose entries need neither to be ordered nor unique. Given the unique and ordered values $\mathbf{v}_{jk}$ and defining $g_{jk,t} = g_{jk}(v_{jk,t})$, the basic idea is to model the vector of functional evaluations $(g_{jk,1}, \ldots, g_{jk,p_{jk}}) \equiv (g_{jk}(v_{jk,1}), \ldots, g_{jk}(v_{jk,p_{jk}}))'$ by viewing them as the realization of a stochastic process that both allows flexibility and also penalizes sharp differences between successive functional evaluations.

Before continuing with details on the stochastic model for $\{g_{jk,t}\}$, we note that unrestricted additive models are identified only up to a constant because the likelihood remains unchanged if $g_{jk}(\cdot)$ and $g_{jh}(\cdot)$, $k \neq h$, in (2.5) 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, Kohn, and Wood (1999) by restricting the functions to equal zero at the first ordered observation (i.e., $g_{jk,1} = 0$), which allows the parametric part of the model to absorb the overall intercept. For the prior distribution of the second state $g_{jk,2}$ of the process, we assume

$$g_{jk,2}|\tau^2_{jk} \sim \mathcal{N}(g_{jk0,2}, \tau^2_{jk} G_{jk0,2}), \tag{2.10}$$

where τ^2_{jk} is a smoothness parameter discussed later.

We model the remaining function evaluations as resulting from the realization of a second-order Markov process. With $h_{jk,t} = v_{jk,t} - v_{jk,t-1}$, the second-order Markov process prior for $g_{jk,t}$, $t = 3, \ldots, p_{jk}$, is

$$\begin{aligned} g_{jk,t} &= \left(1 + \frac{h_{jk,t}}{h_{jk,t-1}}\right) g_{jk,t-1} - \frac{h_{jk,t}}{h_{jk,t-1}} g_{jk,t-2} + u_{jk,t}, \\ u_{jk,t} &\sim \mathcal{N}(0, \tau^2_{jk} h_{jk,t}), \end{aligned} \tag{2.11}$$

where τ^2_{jk} 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. This prior assumes that the variance grows linearly with the distance $h_{jk,t}$, 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).

To see more clearly how this prior specification introduces smoothness, we note several properties of (2.11). First, the expected value of $g_{jk,t}$ given $g_{jk,t-1}$ and $g_{jk,t-2}$ lies on a straight line that passes through $g_{jk,t-1}$ and $g_{jk,t-2}$. Hence, the notion of smoothness that the prior in (2.11) emphasizes is that of local linearity. The possibility of departure from local linearity arises from $u_{jk,t}$, whose variance controls the degree to which deviations are acceptable. The modeling is desirable because, unlike other modeling approaches for the unknown functions, it assumes neither continuous functions nor derivatives.

We include the $\{\tau^2_{jk}\}$ ($j = 1, \ldots, J$, and $k = 1, \ldots, q_j$) in the sampler with prior distributions of the inverse gamma form

$$\tau^2_{jk} \sim \mathcal{IG}(\nu_{jk0}/2, \delta_{jk0}/2). \tag{2.12}$$

The prior specified by (2.10), (2.11), and (2.12) yields a proper and computationally convenient joint prior distribution for the $(p_{jk} - 1)$-vector of unrestricted function evalua-

tions $\mathbf{g}_{jk} = (g_{jk,2}, \ldots, g_{jk,p_{jk}})'$. To see the last point, note that after defining

$$\mathbf{H}_{jk} = \begin{pmatrix} 1 & & & & \\ -\left(1 + \frac{h_{jk,3}}{h_{jk,2}}\right) & 1 & & & \\ \frac{h_{jk,4}}{h_{jk,3}} & -\left(1 + \frac{h_{jk,4}}{h_{jk,3}}\right) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_{jk,p_{jk}}}{h_{jk,p_{jk-1}}} & -\left(1 + \frac{h_{jk,p_{jk}}}{h_{jk,p_{k-1}}}\right) & 1 \end{pmatrix}$$

and $\mathbf{\Sigma}_{jk} = \operatorname{diag}(G_{jk0}, h_{jk,3}, \ldots, h_{jk,p_{jk}})$, the joint distribution of $\mathbf{g}_{jk}|\tau_{jk}^2$ is Gaussian,

$$\mathbf{g}_{jk}|\tau_{jk}^2 \sim \mathcal{N}(\mathbf{g}_{jk0}, \tau_{jk}^2 \mathbf{K}_{jk}^{-1}), \tag{2.13}$$

where $\mathbf{g}_{jk0} = \mathbf{H}_{jk}^{-1}\tilde{\mathbf{g}}_{jk0}$, $\tilde{\mathbf{g}}_{jk0} = (g_{jk0,2}, 0, \ldots, 0)'$ and the *penalty matrix* $\mathbf{K}_{jk} = \mathbf{H}_{jk}'\mathbf{\Sigma}_{jk}^{-1}\mathbf{H}_{jk}$. It is straightforward to verify that because $\mathbf{H}_{jk}$ produces second-order differences when postmultiplied by a vector, its inverse constructs second-order sums. As a result, the vector $\mathbf{H}_{jk}^{-1}\tilde{\mathbf{g}}_{jk0}$ can be constructed by iterating (2.11) in expectation, starting with $g_{jk0,1} = 0$ and $g_{jk0,2}$ from (2.10). It is important to note that $\mathbf{K}_{jk}$ 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 the estimation algorithm, which is an important advantage over modeling approaches for the unknown function that do not result in banded precision or variance matrices. Although bandedness of the precision matrix is a feature of such models as the B-spline approach of Silverman (1985) and the wavelet modeling of Denison et al. (2002), it is not a feature of other modeling approaches, for example, the integrated Wiener process prior of Wahba (1978).

3. ESTIMATION

In this section we first specify an algorithm for the censored selection variable and then turn to the modifications necessary for other data types.

3.1 ALGORITHM FOR CENSORED SELECTION VARIABLE

For notational convenience, we define $\boldsymbol{\theta} \equiv (\boldsymbol{\beta}, \mathbf{\Omega}, \{\mathbf{g}_{jk}\}, \{\tau_{jk}^2\})$, but otherwise continue to use the notation developed in the preceding sections. The posterior distribution for $\boldsymbol{\theta}$ and $\{y_{i4}^*\}_{i \in N_2}$ for the semiparametric model of (2.1)–(2.4) is given by

$$\begin{aligned} &\pi\left(\boldsymbol{\theta}, \{y_{i4}^*\}_{i \in N_2} | \mathbf{y}\right) \\ &\quad \propto \left[\prod_{i \in N_1} f(\mathbf{y}_{i1:4}|\boldsymbol{\theta})\right]\left[\prod_{i \in N_2} f(\mathbf{y}_{i3:4}^*|\boldsymbol{\theta}) I(y_{i4}^* < 0)\right] \mathcal{N}(\boldsymbol{\beta}|\boldsymbol{\beta}_0, \mathbf{B}_0) \\ &\qquad \times \mathcal{IW}(\mathbf{\Omega}|\nu, \mathbf{Q})\left[\prod_{j,k} \mathcal{N}(\mathbf{g}_{jk}|\mathbf{g}_{jk0}, \tau_{jk}^2 \mathbf{K}_{jk}^{-1}) \mathcal{IG}(\tau_{jk}^2|\nu_{jk0}/2, \delta_{jk0}/2)\right]. \end{aligned} \tag{3.1}$$

This posterior distribution can be sampled by Markov chain Monte Carlo (MCMC) methods; see Chib (2001) for a detailed survey. Letting $\boldsymbol{\theta} \setminus \boldsymbol{\theta}_k$ represent the elements of $\boldsymbol{\theta}$ other than $\boldsymbol{\theta}_k$, we summarize the algorithm as follows:

Algorithm 1 (MCMC estimation of nonparametric incidental truncation model)**:**

1. Sample $\boldsymbol{\beta}$ from the distribution $\boldsymbol{\beta}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\beta}$.
2. Sample $\boldsymbol{\Omega}$ from the distribution $\boldsymbol{\Omega}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}$ in a one-block, three-step procedure.
3. For $j = 1, \ldots, J$, $k = 1, \ldots, q_j$, sample $\mathbf{g}_{jk}$ from the distribution $\mathbf{g}_{jk}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \mathbf{g}_{jk}$.
4. For $j = 1, \ldots, J$, $k = 1, \ldots, q_j$, sample τ_{jk}^2 from the distribution $\tau_{jk}^2|\mathbf{y}, \mathbf{g}_{jk}$.
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; that is, we do not augment the selected sample with the data that are not part of the selected sample. This may seem surprising because having the augmented "full" potential sample would reduce the model to a nonparametric seemingly unrelated regression model that could be processed along the lines of Chib and Greenberg (1995), Smith and Kohn (2000), or Holmes, Denison, and Mallick (2002). Specifically, if the missing outcomes are denoted by $\{\mathbf{y}_{i1:2}^\dagger\}$, sampling could proceed recursively by drawing from $[\boldsymbol{\theta}|\mathbf{y}, \mathbf{y}_4^*, \{\mathbf{y}_{i1:2}^\dagger\}]$, $[\mathbf{y}_4^*|\mathbf{y}, \boldsymbol{\theta}, \{\mathbf{y}_{i1:2}^\dagger\}]$, and $[\{\mathbf{y}_{i1:2}^\dagger\}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta}]$, which would be updated by a series of full-conditional draws. In contrast, 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 and storage demands because simulation of the missing $\{\mathbf{y}_{i1:2}^\dagger\}$ is not needed. Second, it significantly improves the mixing of the Markov chain as sampling is not conditional on $\{\mathbf{y}_{i1:2}^\dagger\}$. In Section 5 we present evidence of the improved simulation performance of our method relative to output from a chain that includes $\{\mathbf{y}_{i1:2}^\dagger\}$ and show that the inefficiency factors are much larger when the outcomes that are missing due to the selection mechanism are included. Our results are consistent with the results of Liu (1994) and Liu, Wong, and Kong (1994), who showed that collapsed Gibbs samplers can provide improved simulation performance. These two advantages of the sampler can become important when the proportion of missing outcomes is high or when the number of parameters is 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. It should be noted that, if needed, inference about the missing covariates can be conducted ex post from the probability density function of the missing data conditioned on parameters. We now turn to the details of the sampler.

3.1.1 Sampling $\boldsymbol{\beta}$

The posterior distribution of (3.1) implies $\boldsymbol{\beta}|\mathbf{y}^*, \boldsymbol{\theta} \setminus \boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}, \mathbf{B})$, where

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

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

where $\mathbf{J}$ was defined in (2.8). This step proceeds without the unobserved $\mathbf{y}_{i1:2}$ by computing the conditional mean and covariance of $(\beta_1', \beta_2')'$ from the observations in N_1 and the conditional mean and covariance of $(\beta_3', \beta_4')'$ from the observations in both N_1 and N_2. The matrix $\mathbf{J}$ selects β_3 and β_4 from β to include in the computations the observations in N_2. The partitioning of $\boldsymbol{\Omega}$ is necessary because the observations in N_2 are modeled by the third and fourth equations only.

3.1.2 Sampling $\boldsymbol{\Omega}$

Our decision not to sample the missing $\mathbf{y}_{i1:2}$ also has implications for the way in which $\boldsymbol{\Omega}$ is sampled. In particular, the conditional distribution $\boldsymbol{\Omega}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}$ is not inverse Wishart, because of the different forms of the complete-data likelihood in N_1 and N_2. It is nevertheless possible to derive and sample the distributions $\boldsymbol{\Omega}_{22}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{11\cdot2}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}_{11\cdot2}$, and $\mathbf{B}_{21}|\mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\Omega}_{11\cdot2}$, where

$$\boldsymbol{\Omega}_{11\cdot2} = \boldsymbol{\Omega}_{11} - \boldsymbol{\Omega}_{12}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21},$$
$$\mathbf{B}_{21} = \boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\Omega}_{21},$$

from which $\boldsymbol{\Omega}$ can be recovered. To see the form of these three conditional distributions, let

$$\boldsymbol{\eta}_{i1:4} = \mathbf{y}_{i1:4} - \mathbf{g}_{i1:4} - \mathbf{X}_{i1:4}\boldsymbol{\beta}, \qquad i \in N_1,$$
$$\boldsymbol{\eta}_{i3:4}^* = \mathbf{y}_{i3:4}^* - \mathbf{g}_{i3:4} - \mathbf{X}_{i3:4}\mathbf{J}'\boldsymbol{\beta}, \qquad i \in N_2.$$

The complete-data likelihood is then

$$\prod_{i\in N_1} f(\boldsymbol{\eta}_{i1:4}|\boldsymbol{\theta}) \prod_{i\in N_2} f(\boldsymbol{\eta}_{i3:4}^*|\boldsymbol{\theta})$$
$$\propto |\boldsymbol{\Omega}|^{-n_1/2} \exp\left[-\frac{1}{2}\sum_{i\in N_1} \boldsymbol{\eta}_{i1:4}'\boldsymbol{\Omega}^{-1}\boldsymbol{\eta}_{i1:4}\right]$$
$$\times |\boldsymbol{\Omega}_{22}|^{-n_2/2} \exp\left[-\frac{1}{2}\sum_{i\in N_2} \boldsymbol{\eta}_{i3:4}^{*\prime}\boldsymbol{\Omega}_{22}^{-1}\boldsymbol{\eta}_{i3:4}^*\right].$$

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

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

we find the posterior full-conditional distribution of $\boldsymbol{\Omega}$:

$$\pi(\boldsymbol{\Omega}|\mathbf{y}, \mathbf{y}_4^*, \theta \setminus \boldsymbol{\Omega})$$
$$\propto |\boldsymbol{\Omega}|^{-(\nu+n_1+J_1+J_2+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})\right]$$
$$\times |\boldsymbol{\Omega}_{22}|^{-n_2/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Omega}_{22}^{-1}\sum_{i\in N_2} \boldsymbol{\eta}_{i3:4}^*\boldsymbol{\eta}_{i3:4}^{*\prime}\right)\right],$$

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

Making the change of variables from $\boldsymbol{\Omega}$ to $(\boldsymbol{\Omega}_{22}, \boldsymbol{\Omega}_{11\cdot 2}, \mathbf{B}_{21})$, with Jacobian $|\boldsymbol{\Omega}_{22}|^{J_1}$, we obtain the posterior distribution

$$\begin{aligned}
&\pi(\boldsymbol{\Omega}_{22}, \boldsymbol{\Omega}_{11\cdot 2}, \mathbf{B}_{21} | \mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}) \\
&\quad \propto |\boldsymbol{\Omega}_{11\cdot 2}|^{-(\nu+n_1+J_1+J_2+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R})\right] \\
&\qquad \times |\boldsymbol{\Omega}_{22}|^{-(\nu+n-J_1+J_2+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Omega}_{22}^{-1}\sum_{i\in N_2}\boldsymbol{\eta}_{i3:4}^*\boldsymbol{\eta}_{i3;4}^{*\prime}\right)\right],
\end{aligned}$$

where we use $|\boldsymbol{\Omega}| = |\boldsymbol{\Omega}_{11\cdot 2}||\boldsymbol{\Omega}_{22}|$. By the partitioned inverse theorem

$$\boldsymbol{\Omega}^{-1} = \begin{pmatrix} \boldsymbol{\Omega}_{11\cdot 2}^{-1} & -\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{B}_{21}' \\ -\mathbf{B}_{21}\boldsymbol{\Omega}_{11\cdot 2}^{-1} & \boldsymbol{\Omega}_{22}^{-1} + \mathbf{B}_{21}\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{B}_{21}' \end{pmatrix},$$

we are able to simplify the trace as

$$\begin{aligned}
\operatorname{tr}(\boldsymbol{\Omega}^{-1}\mathbf{R}) &= \operatorname{tr}\big([\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{R}_{11} - \boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{B}_{21}'\mathbf{R}_{21} - \mathbf{B}_{21}\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{R}_{12} \\
&\qquad + (\boldsymbol{\Omega}_{22}^{-1} + \mathbf{B}_{21}\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{B}_{21}')\mathbf{R}_{22}]\big) \\
&= \operatorname{tr}(\boldsymbol{\Omega}_{11\cdot 2}^{-1}[\mathbf{R}_{11} + \mathbf{B}_{21}'\mathbf{R}_{22}\mathbf{B}_{21} - \mathbf{B}_{21}'\mathbf{R}_{21} - \mathbf{R}_{12}\mathbf{B}_{21}]) + \operatorname{tr}(\boldsymbol{\Omega}_{22}^{-1}\mathbf{R}_{22}) \\
&= \operatorname{tr}\big(\boldsymbol{\Omega}_{11\cdot 2}^{-1}[(\mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}) + (\mathbf{B}_{21} - \mathbf{R}_{22}^{-1}\mathbf{R}_{21})'\mathbf{R}_{22}(\mathbf{B}_{21} - \mathbf{R}_{22}^{-1}\mathbf{R}_{21})]\big) \\
&\qquad + \operatorname{tr}(\boldsymbol{\Omega}_{22}^{-1}\mathbf{R}_{22}),
\end{aligned}$$

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

$$\begin{aligned}
&\pi(\boldsymbol{\Omega}_{22}, \boldsymbol{\Omega}_{11\cdot 2}, \mathbf{B}_{21} | \mathbf{y}^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega}) \\
&\quad \propto |\boldsymbol{\Omega}_{11\cdot 2}|^{-(\nu+n_1+J_1+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}_{11\cdot 2}^{-1}\mathbf{R}_{11\cdot 2})\right] \\
&\qquad \times |\boldsymbol{\Omega}_{11\cdot 2}|^{-J_2/2} \exp\left[-\frac{1}{2}\operatorname{tr}(\boldsymbol{\Omega}_{11\cdot 2}^{-1}(\mathbf{B}_{21} - \mathbf{R}_{22}^{-1}\mathbf{R}_{21})'\mathbf{R}_{22}(\mathbf{B}_{21} - \mathbf{R}_{22}^{-1}\mathbf{R}_{21}))\right] \\
&\qquad \times |\boldsymbol{\Omega}_{22}|^{-(\nu+n-J_1+J_2+1)/2} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\boldsymbol{\Omega}_{22}^{-1}\left[\mathbf{R}_{22} + \sum_{N_2}\boldsymbol{\eta}_{i3:4}\boldsymbol{\eta}_{i3:4}'\right]\right)\right],
\end{aligned}$$

where $\mathbf{R}_{11\cdot 2} = \mathbf{R}_{11} - \mathbf{R}_{12}\mathbf{R}_{22}^{-1}\mathbf{R}_{21}$. From here we conclude that

$$\begin{aligned}
\boldsymbol{\Omega}_{22} | \mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega} &\sim \mathcal{IW}\left(\nu - J_1 + n_1 + n_2, \mathbf{Q}_{22} + \sum_{N_1,N_2}\boldsymbol{\eta}_{i3:4}^*\boldsymbol{\eta}_{i3:4}^{*\prime}\right), \\
\boldsymbol{\Omega}_{11\cdot 2} | \mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\theta} \setminus \boldsymbol{\Omega} &\sim \mathcal{IW}(\nu + n_1, \mathbf{R}_{11\cdot 2}), \\
\mathbf{B}_{21} | \mathbf{y}, \mathbf{y}_4^*, \boldsymbol{\Omega}_{11\cdot 2} &\sim \mathcal{MN}_{J_2\times J_1}(\mathbf{R}_{22}^{-1}\mathbf{R}_{21}, \boldsymbol{\Omega}_{11\cdot 2} \otimes \mathbf{R}_{22}^{-1}).
\end{aligned}$$

The sampling can thus proceed from the full conditional densities of $\boldsymbol{\Omega}_{22}$, $\boldsymbol{\Omega}_{11\cdot 2}$, and $\mathbf{B}_{21}$, from which $\boldsymbol{\Omega}$ can be recovered.

3.1.3 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. Because the sampling of each function is conditional on all parameters and other functions, in sampling the kth function we can focus only on the equation containing that function. Letting

$$y_{ij}^* = \begin{cases} y_{ij}, & \text{if } j = 1, 2, 3 \\ y_{i4}^*, & \text{if } j = 4, \end{cases}$$

to isolate the kth function in equation j, we define $\xi_{ijk} \equiv y_{ij}^* - x_{ij}'\beta_j - \sum_{h \neq k} g_{jh}(w_{ijh}) - E(\varepsilon_{ij}|\varepsilon_{i\setminus j})$, so that

$$\xi_{ijk} = g_{jk}(w_{ijk}) + \hat{\varepsilon}_{ijk},$$

where $\hat{\varepsilon}_{ijk} \overset{\text{ind}}{\sim} N(0, \text{var}(\varepsilon_{ij}|\varepsilon_{i\setminus j}))$. Note that we can condition on $\varepsilon_{i\setminus j}$ because we are conditioning upon $\boldsymbol{\theta} \setminus g_{jk}$ in this computation. Stacking over the n_1 observations if equation j is part of the selected sample ($j \leq J_1$) or over the n observations if equation j is always observed ($J_1 < j \leq J$), we can write

$$\boldsymbol{\xi}_{jk} = \mathbf{P}_{jk}\mathbf{g}_{jk} + \hat{\varepsilon}_{jk}, \tag{3.2}$$

where $\mathbf{g}_{jk}$ is the $(p_{jk} - 1)$-vector of unrestricted function evaluations defined on $\tilde{\mathbf{v}}_{jk} = (v_{jk,2}, \ldots, v_{jk,p_{jk}})$ as discussed in Section 2.2. The matrix $\mathbf{P}_{jk}$ is an $n \times (p_{jk} - 1)$ or $n_1 \times (p_{jk} - 1)$ *incidence matrix* with entries $\mathbf{P}_{jk}(h, l) = 1$ if $w_{hjk} = v_{jk,l+1}$ and 0 otherwise, which establishes the correspondence between $\mathbf{w}_{jk}$ and $\tilde{\mathbf{v}}_{jk}$ over which the unrestricted function evaluations are defined. Because the rows of $\mathbf{P}_{jk}$ for which $w_{hjk} = v_{jk,1}$ contain only zeros, whereas all other rows contain a single 1, row i of the product $\mathbf{P}_{jk}\mathbf{g}_{jk}$ is $g_{jk}(w_{ijk})$.

A closer look at (3.2) can help demystify the nature of nonparametric modeling. It is instructive to note that, when written in this way, the model specifies a dummy variable at each unique covariate observation, with $g_{jk}(v_{jk,1})$ serving as the omitted category. In this model, the $\mathbf{g}_{jk}$ can be interpreted as the parameters on which the prior in Section 2.2 imposes smoothness without ruling out any values they may take, which explains why we consider $g_{jk}(\cdot)$ to be a nonparametric function.

It now follows from standard calculations that

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

where

$$\hat{\mathbf{G}}_{jk} = (\tau_{jk}^{-2}\mathbf{K}_{jk} + \mathbf{P}_{jk}'\mathbf{V}_{jk}^{-1}\mathbf{P}_{jk})^{-1},$$
$$\hat{\mathbf{g}}_{jk} = \hat{\mathbf{G}}_{jk}(\tau_{jk}^{-2}\mathbf{K}_{jk}\mathbf{g}_{jk0} + \mathbf{P}_{jk}'\mathbf{V}_{jk}^{-1}\xi_{jk}),$$

and $\mathbf{V}_{jk}$ is a diagonal matrix with entries equal to $\text{var}(\varepsilon_{ij}|\varepsilon_{i\setminus j})$, which introduces heteroscedasticity into the sampling of the unknown functions.

In sampling $\mathbf{g}_{jk}$, one should note that $\mathbf{P}_{jk}'\mathbf{V}_{jk}^{-1}\mathbf{P}_{jk}$ is a diagonal matrix with the tth diagonal entry equal to the number of values in $\mathbf{w}_{jk}$ corresponding to the tth entry in $\tilde{\mathbf{v}}_{jk}$ divided by $\text{var}(\varepsilon_{ij}|\varepsilon_{i\setminus j})$. Because $\mathbf{K}_{jk}$ and $\mathbf{P}_{jk}'\mathbf{V}_{jk}^{-1}\mathbf{P}_{jk}$ are banded, $\hat{\mathbf{G}}_{jk}^{-1}$ is banded as well, and sampling of $\mathbf{g}_{jk}$ does not require an inversion to obtain $\hat{\mathbf{G}}_{jk}$ and $\hat{\mathbf{g}}_{jk}$. Instead, the mean $\hat{\mathbf{g}}_{jk}$ can be found by solving $\hat{\mathbf{G}}_{jk}^{-1}\hat{\mathbf{g}}_{jk} = \tau_{jk}^{-2}\mathbf{K}_{jk}\mathbf{g}_{jk0} + \mathbf{P}_{jk}'\mathbf{V}_{jk}^{-1}\xi_{jk}$ by back substitution

in $O(n)$ operations. A random draw from $\mathcal{N}(\hat{\mathbf{g}}_{jk}, \hat{\mathbf{G}}_{jk})$ can then be efficiently obtained by sampling $\nu \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and solving $\mathbf{C}\mathbf{z} = \nu$ for $\mathbf{z}$ by back substitution, where $\mathbf{C}$ is the Cholesky decomposition of $\hat{\mathbf{G}}_{jk}^{-1}$ and is also banded. It follows that $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \hat{\mathbf{G}}_{jk})$, so that adding the mean $\hat{\mathbf{g}}_{jk}$ to $\mathbf{z}$ gives a draw $\mathbf{g}_{jk} \sim \mathcal{N}(\hat{\mathbf{g}}_{jk}, \hat{\mathbf{G}}_{jk})$.

3.1.4 Sampling τ_{jk}^2

The smoothness parameters τ_{jk}^2 for each unknown function ($j = 1, \ldots, J, k = 1, \ldots, q_j$) are sampled from

$$\tau_{jk}^2 | \boldsymbol{\theta} \setminus \tau_{jk}^2 \sim \mathcal{IG}\left(\frac{\nu_{jk0} + p_{jk} - 1}{2}, \frac{\delta_{jk0} + (\mathbf{g}_{jk} - \mathbf{g}_{jk0})' \mathbf{K}_{jk} (\mathbf{g}_{jk} - \mathbf{g}_{jk0})}{2} \right).$$

3.1.5 Sampling y_{i4}^*

Following Chib (1992), this full conditional density is seen to be truncated normal:

$$y_{i4}^* | \mathbf{y}, \boldsymbol{\theta} \sim \mathcal{TN}_{(-\infty,0)}\big(\mathbf{x}_{i4}' \boldsymbol{\beta}_4 + g_4(\mathbf{w}_{i4}) + E(\varepsilon_{i4} | \varepsilon_{i\setminus 4}), \operatorname{var}(\varepsilon_{i4} | \varepsilon_{i\setminus 4})\big), \qquad i \in N_2.$$

3.2 MODIFICATIONS FOR MULTIPLE QUALITATIVE VARIABLES

If there is more than one qualitative variable in the model—for example, the response variable or one or more of the endogenous variables is binary—three modifications to the basic scheme are required, following the framework of Albert and Chib (1993). First, y_{ij}^* is substituted for y_{ij} in the specification of the likelihood function, analogously to the use of y_{i4}^* in the selection equation. Second, y_{ij}^* is added to the sampler and sampled from an appropriately truncated distribution. Third, the variances of any binary or ordinal variables are set to 1. 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 to the algorithm to reflect the unit-variance constraints. When only one variable is binary or ordinal and that variable is always observed, the method for sampling $\boldsymbol{\Omega}$ presented in Section 3.1 may be utilized repeatedly to produce a draw for $\boldsymbol{\Omega}$. For instance, if y_{i4} is binary, rather than censored, the matrix $\boldsymbol{\Omega}_{22}$ in (2.7) is

$$\boldsymbol{\Omega}_{22} = \begin{pmatrix} \omega_{33} & \omega_{34} \\ \omega_{34} & 1 \end{pmatrix},$$

so that the full conditional distributions for ω_{34} and $\omega_{33\cdot 4} = \omega_{33} - \omega_{34}^2$ are normal and inverse Wishart (or inverse gamma when $\omega_{33\cdot 4}$ is univariate), respectively. Notice that if $\omega_{33} = 1$ and ω_{44} is free, an obvious reindexing would again permit Gibbs sampling from inverse Wishart (or inverse gamma) and Gaussian distributions. Once $\boldsymbol{\Omega}_{22}$ is obtained, the rest of $\boldsymbol{\Omega}$ can be simulated in a straightforward way with the sampler in Section 3.1. A sampler for $\boldsymbol{\Omega}$ in which one of the variances is restricted to unity was presented in Munkin and Trivedi (2003) in a setting with endogeneity but without incidental truncation.

When more than one variable is qualitative or when a unit restriction appears in $\boldsymbol{\Omega}_{11}$, that is, there is a qualitative variable in $y_{i1:2}$, the above derivations leading to normal and inverse

Wishart sampling steps cannot be applied. Instead, the possibly multiple unit-variance and positive definiteness restrictions on $\boldsymbol{\Omega}$ require a Metropolis–Hastings algorithm as in Chib and Greenberg (1998).

It is difficult to offer guidance on how such complications affect computational time. The key burden is in the sampling of the free elements of the restricted $\boldsymbol{\Omega}$. Although there are no costs of sampling such fixed elements of $\boldsymbol{\Omega}$ as unit variances or zero correlations, and the computational costs can be quite small if $\boldsymbol{\Omega}$ is well-structured and parameterized parsimoniously (e.g., equicorrelated or Toeplitz), sampling of the free elements requires an M–H step for which the implementation cost depends on whether tailored approximations to the full-conditional for $\boldsymbol{\Omega}$ can be easily obtained by Taylor series methods or otherwise.

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

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

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

$$m(\mathbf{y}) = \frac{f(\mathbf{y}|\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_{jk}^*\}, \{\tau_{jk}^{2*}\})\pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_{jk}^*\}, \{\tau_{jk}^{2*}\})}{\pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_{jk}^*\}, \{\tau_{jk}^{2*}\}|\mathbf{y})}, \tag{4.1}$$

where $\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_{jk}^*\}$, and $\{\tau_{jk}^{2*}\}$ are fixed at high-density values. The posterior ordinate in the denominator of (4.1) must be estimated, but the likelihood and the prior ordinates in the numerator are directly available. To estimate the denominator we use the decomposition

$$\begin{aligned}\pi(\boldsymbol{\beta}^*, \boldsymbol{\Omega}^*, \{\mathbf{g}_{jk}^*\}, \{\tau_{jk}^{2*}\}|\mathbf{y}) &= \pi(\boldsymbol{\Omega}^*, \{\tau_{jk}^{2*}\}|\mathbf{y})\pi(\boldsymbol{\beta}^*|\mathbf{y}, \boldsymbol{\Omega}^*, \{\tau_{jk}^{2*}\}) \\ &\quad \times \prod_{\mathcal{I}_{mn}=1}^{q} \pi\big(\mathbf{g}_{\mathcal{I}_{mn}}^*|\mathbf{y}, \boldsymbol{\Omega}^*, \boldsymbol{\beta}^*, \{\tau_{jk}^{2*}\}, \{\mathbf{g}_{\mathcal{I}_{jk}}^*\}_{\mathcal{I}_{jk}<\mathcal{I}_{mn}}\big),\end{aligned}$$

where $\mathcal{I}_{jk}$ denotes a particular indexing of the q functions in the set $\{\mathbf{g}_{jk}\}$, so that it is not necessary to estimate the ordinates for each function in the order in which the functions appear in the model. The terms in the product are estimated by Rao–Blackwellization by averaging the full conditional densities of 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 reduced runs in which the parameters whose ordinates have already been obtained are held fixed and sampling is over the remaining elements of $\boldsymbol{\theta}$ and the latent data $\{y_{i4}^*\}$:

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

for draws $\{\boldsymbol{\theta}_j^{(g)}\}_{j>s} \sim \pi(\{\boldsymbol{\theta}_j^{(g)}\}_{j>s}|\mathbf{y}, \{\boldsymbol{\theta}_j^*\}_{j<s})$, $g = 1, \ldots, G$.

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 $\{\mathbf{g}_{jk}\}$, 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 toward 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 because the sampling of the unknown functions is $O(n)$. Finally, the ordinate $\pi(\boldsymbol{\Omega}^*, \{\tau_{jk}^{2*}\}|\mathbf{y})$ is estimated jointly, rather than in individual reduced runs, because the parameters $\boldsymbol{\Omega}$ and $\{\tau_{jk}^2\}$ are conditionally independent given $\boldsymbol{\beta}$ and $\{\mathbf{g}_{jk}\}$. This observation can significantly reduce the computations.

5. SIMULATION STUDY

To study the effectiveness of the sampler in estimating nonparametric functions we simulate data from a three-equation version of the model in (2.1)–(2.4) that is partly motivated by our subsequent application to log-wages considered in Section 6. In our simulation study each equation contains an intercept and two additive functions:

$$y_{i1} = \beta_1 + g_{11}(y_{i2}) + g_{12}(w_{i11}) + \varepsilon_{i1}, \tag{5.1}$$

$$y_{i2} = \beta_2 + g_{21}(w_{i21}) + g_{22}(w_{i22}) + \varepsilon_{i2}, \tag{5.2}$$

$$y_{i3}^* = \beta_3 + g_{31}(w_{i31}) + g_{32}(w_{i32}) + \varepsilon_{i3}. \tag{5.3}$$

The nonparametric functions, graphed in Figure 1, have previously appeared in the literature: $g_{11}(y_{i2}) = 2\Phi(v) - 1$, 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 $m = 51$ equally spaced points on the support of each function, and $g_{11}(y_{i2})$ is evaluated at all points generated by the random realizations of y_{i2}. The data are generated with the equicorrelated covariance matrix $\boldsymbol{\Omega} = (0.15\mathbf{I}_4 + 0.1\mathbf{i}\mathbf{i}')$, which implies a relatively high correlation of 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 equal (e.g., the functional form, the sample size, and the number of observations at each design point), the functions tend to be estimated more precisely as the signal-to-noise ratio is increased (cf. Wood et al. 2002). This study covers a reasonably "strong noise" scenario, and the performance of the

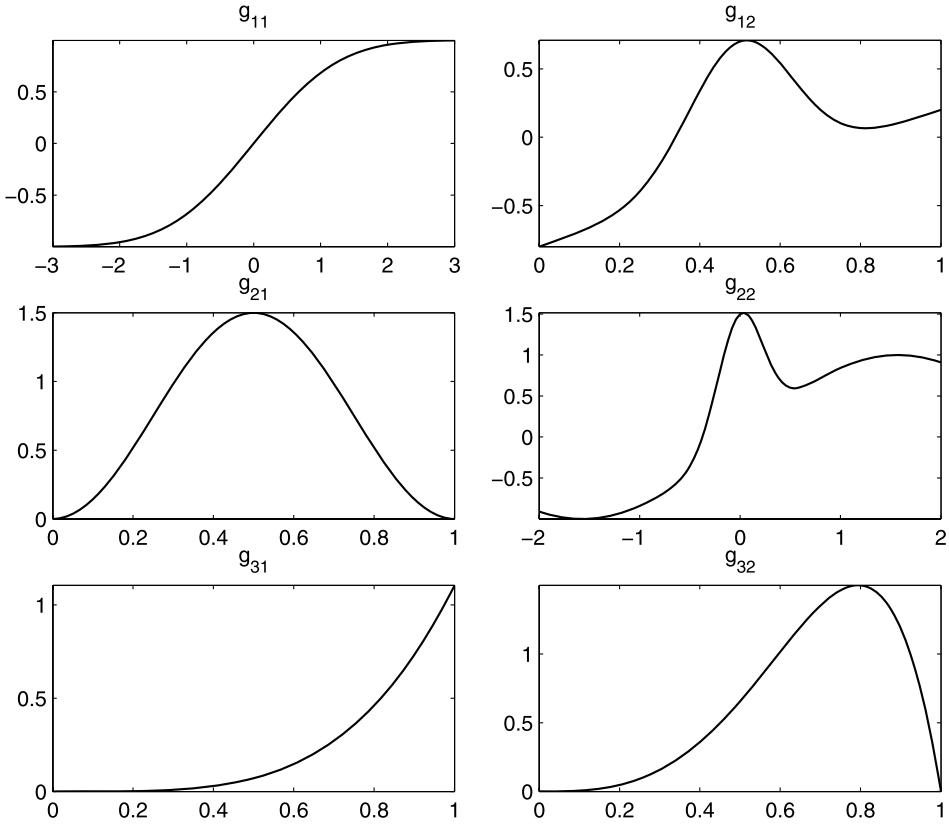

Figure 1. Functions used in the simulation study.

techniques is likely to improve as the sample size is increased, more points are introduced at each design point, or the error variances are decreased.

To simplify the discussion of prior distributions, we denote the nonlinear functions by g_k, $k = 1, \ldots, 6$, in consecutive order over the three equations. Comparable priors are set for the equations: $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 5 \times \mathbf{I})$, $\boldsymbol{\Omega} \sim \mathcal{IW}(J + 4, 1.2 \times \mathbf{I}_J)$, $g_k|\tau_k^2 \sim \mathcal{N}(0, \tau_k^2/E(\tau_k^2))$, $\tau_1^2 \sim \mathcal{IG}(6, 0.0004)$, and $\tau_k^2 \sim \mathcal{IG}(6, 0.04)$ for $k = 2, \ldots, 6$; these priors imply that $E(\boldsymbol{\Omega}) = 0.4 \times \mathbf{I}$, $\mathrm{SD}(\mathrm{diag}(\boldsymbol{\Omega})) = 0.57 \times \mathbf{1}$, $E(\tau_1^2) = \mathrm{SD}(\tau_1^2) = 0.0001$, $E(\tau_k^2) = \mathrm{SD}(\tau_k^2) = 0.01$ for $k = 2, \ldots, 6$, and $E_{\tau_k^2}(\mathrm{var}(g_k|\tau_k^2)) = 1$, for $k = 1, \ldots, 6$. A tighter prior on the smoothness parameter is assumed for the first function because most of the generated y_{i2}

Table 1. Signal-to-noise ratios for the functions in the simulation study.

	Functions					
	g_{11}	g_{12}	g_{21}	g_{22}	g_{31}	g_{32}
$\mathrm{Range}(g_k^j)/\mathrm{SD}(\varepsilon_{ij})$	4.0	3.0	3.0	5.0	2.2	3.0

NOTE: The ratio for $g_{11}(\cdot)$ is an upper bound because the y_{i2} are randomly generated and subsequently censored, thus unlikely to fill the entire range.

appear around the middle of the support, where the function is approximately linear. Of course, the data play a role along with the prior in determining the posterior distribution of τ_1^2. As is well known in the literature, high values of τ^2 lead to undersmoothing as the function becomes less smooth and tries to interpolate the observations, whereas low values lead to smoother functions. Values of τ^2 that are too high or too low yield poorer approximations that tend to improve as more data become available. An example of the role of the smoothness parameter in over- and undersmoothing was presented in Chib and Jeliazkov (2006).

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: under the simulation design, the selected sample n_1 comprises approximately 85% of the potential sample n and varies from a low of 81.6% to a high of 86.6%. In comparison, our log-wage application in Section 6 has $n = 753$ and $n_1 = 428$, whereas in our simulated data, when $n = 500$, n_1 varies from 408 to 433. An important feature of the real data application compared to our simulation study is the presence of repeated values in the real data, which aids estimation. In our simulation study, most functions are specified to have $p = 51$ knots, except for $g_{11}(\cdot)$, which has n_1 knots with only one observation per knot, so that, in the first equation, the estimated $\mathbf{g}_{11}$ and $\mathbf{g}_{12}$ jointly contain more elements than the selected sample size n_1. Although the remaining equations require the estimation of lower-dimensional objects, the simulated data require more parameters than the real data application to log-wages, where all functions have fewer than 50 knots (ranging from 18 to 45), allowing more data to be available at each knot, which serves to estimate the functions more reliably than in the simulated data.

The posterior mean estimates $\hat{\mathbf{g}}_{jk} = E(\mathbf{g}_{jk}(v)|\mathbf{y})$ are found from MCMC runs of length 15,000 following 2500 burn-in cycles. In the current context, the computational cost per 1000 MCMC draws is approximately 19 seconds when $n = 500$ and grows linearly with the sample size (approximately 38 seconds when $n = 1000$ and approximately 77 seconds when $n = 2000$), as is to be expected of an $O(n)$ algorithm. We gauge the performance of the method in fitting these functions by the root mean squared error, $\text{RMSE}_{jk} = \sqrt{\frac{1}{m}\sum_{i=1}^{m}\{\hat{g}_{jk}(v_{ijk}) - g_{jk}(v_{ijk})\}^2}$, where the true functions are shifted so as to satisfy the identification constraints of the model. Boxplots of the RMSE_k 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. The function $g_{11}(y_{i2})$ 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, whereas 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 (5.1) not only is related 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 from the description of the sampler for $\mathbf{g}$, especially (3.2), which shows that the errors in the endogenous covariate equation determine both the endogenous covariate and the conditional mean used in sam-

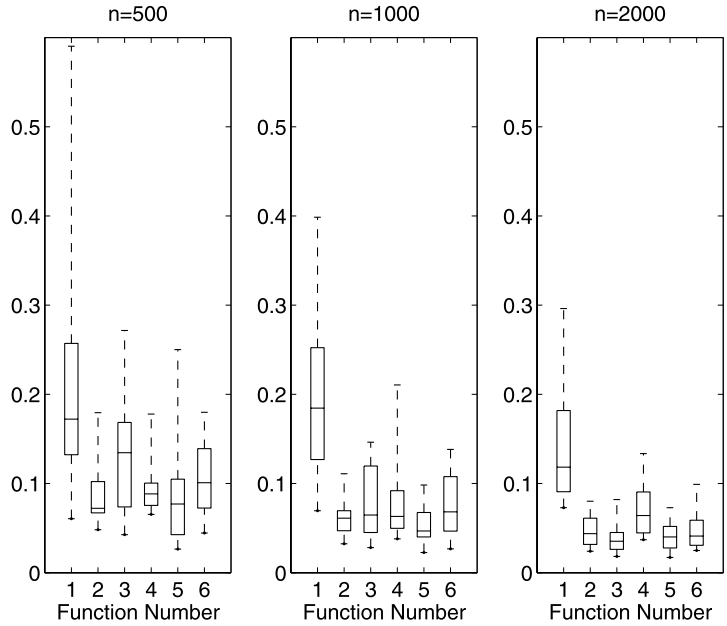

Figure 2. Boxplots of root mean squared errors of function estimates in the simulation study.

pling $g_{11}(\cdot)$ 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 model. Plots of the true and estimated functions for $n = 1000$ are shown 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 $1 + 2\sum_{l=1}^{L} \rho_k(l)(1 - l/L)$, where $\rho_k(l)$ is the sample autocorrelation at lag l for the kth parameter in the sampling and the summation is 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 3.1, suggests several conclusions. First, although the inefficiency factors for $\boldsymbol{\beta}$ 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 2007). For this reason, we suggest that a longer MCMC chain may be required for more accurate estimation of $\boldsymbol{\beta}$, 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, because the estimates of the $\{\tau_{jk}^2\}$ depend on how well the corresponding functions are sampled, they depend on the sample size in a predictable way: as the sample size grows and the $\{\mathbf{g}_{jk}\}$ are estimated better, so are the corresponding $\{\tau_{jk}^2\}$.

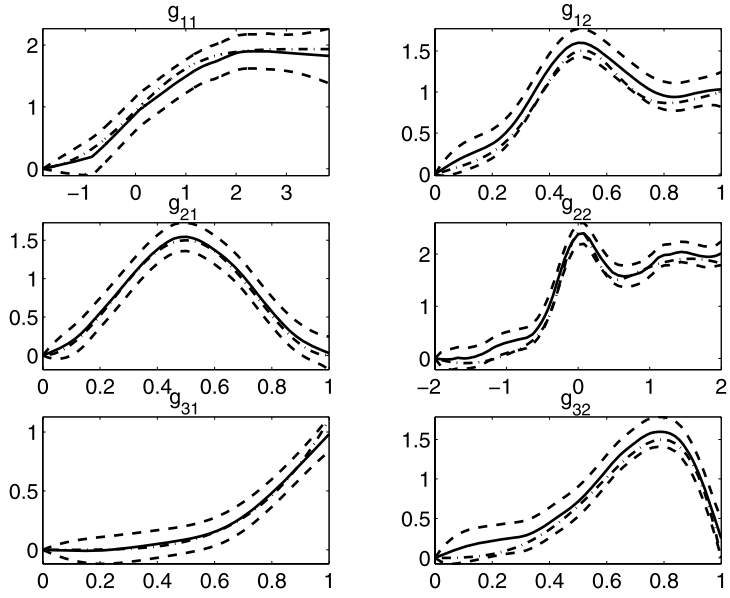

Figure 3. An example of function estimates in the simulation study: true functions (dot–dashes), estimated functions (solid lines), and pointwise confidence bands (dashes).

We conclude the discussion of the simulated data by revisiting the discussion, below Algorithm 1 of Section 3.1, of the advantages of not involving the outcomes that are missing due to the selection mechanism. The inefficiency factors for our algorithm, which does

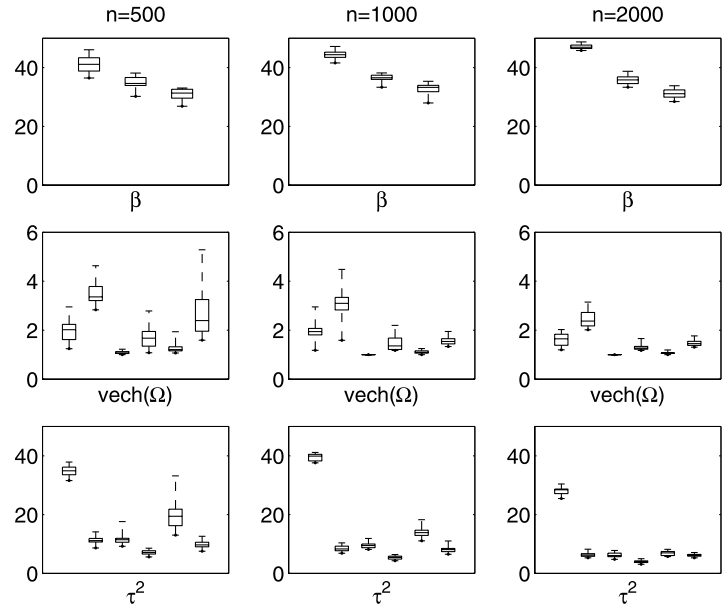

Figure 4. Inefficiency factors for the parameters of the model in the simulation study.

not augment the sampler with the missing outcomes, are summarized in Figure 4. We next compare the performance of our algorithm to that of an algorithm that includes the missing outcomes. Specifically, samples are obtained from an algorithm in which $\{\mathbf{g}_{jk}\}$, $\{\tau_{jk}^2\}$, $\boldsymbol{\beta}$, and $\{y_{i4}^*\}$ are sampled as before (without involving the missing data for the incidentally truncated outcomes). To sample $\boldsymbol{\Omega}$, however, we first augment $\boldsymbol{\eta}_{i2:3}^*$ for $i \in N_2$ with the missing $\boldsymbol{\eta}_{i1}$ (see the sampler for $\boldsymbol{\Omega}$ in Section 3.1, but keep in mind that here we are fitting a three-equation system), which are drawn conditional on $\boldsymbol{\eta}_{i2:3}^*$ and all other parameters and data, including $\boldsymbol{\Omega}$. After the data are "balanced" in this way, $\boldsymbol{\Omega}$ can be sampled directly from an inverse Wishart distribution.

The results from the sampler augmented with the outcomes that are missing due to the selection mechanism are presented in Figure 5. A comparison of Figures 4 and 5 shows that inclusion of the missing outcomes results in drastic deterioration of the inefficiency factors. This result is most obvious in the inefficiency factors for the first equation, where augmentation is performed to "balance" the sample, and all sample sizes: for a sample of 500, the inefficiency factor without latent data for β_1 has a median of about 40, whereas its median is close to 200 when latent data are included. As another example, for $n = 2000$, the median is about 50 without the missing outcomes and 400 with the missing outcomes included in the sampling. Again, with $n = 2000$, the median inefficiency factor for ω_{12} is about 2.5 without the missing outcomes and about 10 with them. We make several observations regarding these results. First, the sampler we introduced in Section 3 dominates a sampler augmented with the missing outcomes on all margins: the additional computational and storage demands of introducing the missing outcomes related to augmenting the sample with missing data do not pay off as the inefficiency factors are significantly higher.

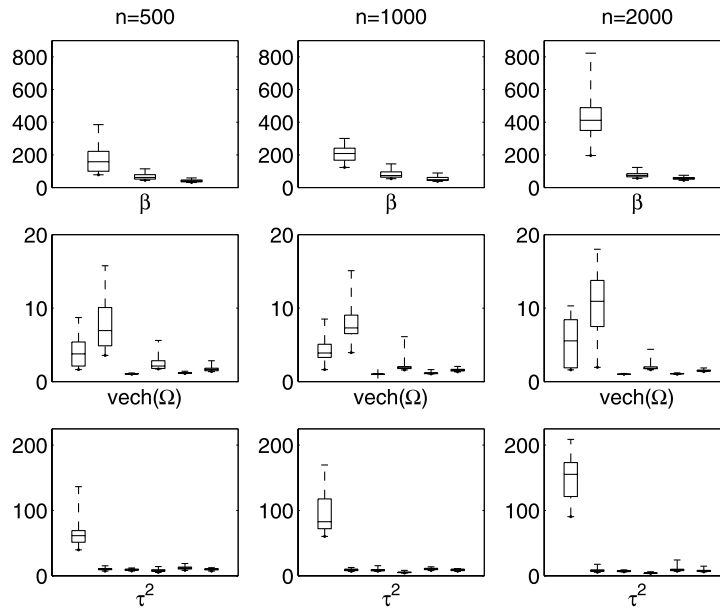

Figure 5. Inefficiency factors for the parameters of the model in the simulation study; latent unobserved data included in the sampler.

Second, Figure 5 reveals that the inefficiency factors actually rise in larger samples. This may seem surprising because we expect samplers to improve as more data become available. But an important feature of this model is that larger potential samples imply both larger selected samples and larger amounts of missing data to simulate, which slows the convergence of the Markov chain. Finally, note that we generate auxiliary data only in the sampling of $\mathbf{\Omega}$, whereas all other parameters are sampled marginally of it. If one were to try an even more "naive" approach, where the missing data are involved in all steps of the sampler, the results would be even worse than those we have reported for the reasons we have mentioned. These conclusions are in close agreement with the results in Liu (1994) and Liu, Wong and Kong (1994), who emphasized the importance of constructing Markov chains that sample parameters jointly instead of conditionally on each other.

6. ESTIMATING WOMEN'S WAGES

We apply the techniques of this article 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 United States in the postwar period. Goldin (1989) reported a 7-fold increase in participation of married women since the 1920s, and Heckman (1993) underscored 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 concerns. Endogeneity may be a concern for the level of education, a covariate that affects wages, because education is likely to be affected by such unobserved variables as motivation, work ethic, perseverance, and intelligence, that are also determinants of wage rates. The problem of sample selection 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 the nonparametric functions allow us to explore the presence of nonlinearities in the effects of some of the covariates. Such nonlinearities have been modeled by the inclusion of quadratic terms for certain variables; see Mroz (1987) and Wooldridge (2002). We extend these results by estimating a semiparametric specification and comparing it to a number of alternative models.

The dataset, which is available on the *JCGS* website, is from Mroz (1987). The potential sample consists of 753 married women, 428 of whom are employed and therefore in the selected sample. The variables in the dataset are summarized in Table 2.

We specify the econometric model as

$$y_{i1} = \mathbf{x}_{i1}'\boldsymbol{\beta}_1 + g_{11}(y_{i2}) + g_{12}(w_{i11}) + \varepsilon_{i1}, \tag{6.1}$$

$$y_{i2} = \mathbf{x}_{i2}'\boldsymbol{\beta}_2 + g_{21}(w_{i21}) + g_{22}(w_{i22}) + \varepsilon_{i2}, \tag{6.2}$$

$$y_{i3}^* = \mathbf{x}_{i3}'\boldsymbol{\beta}_3 + g_{31}(w_{i31}) + g_{32}(w_{i32}) + \varepsilon_{i3}, \tag{6.3}$$

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

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.

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

Mroz (1987) and Wooldridge (2002), we let

$$\mathbf{y}_i = (y_{i1}, y_{i2}, y_{i3})' = (\ln(\text{WAGE}_i), \text{EDU}_i, \sqrt{\text{HRS}_i})', \qquad \mathbf{x}_{i1} = 1,$$

$$\mathbf{x}_{i2} = \mathbf{x}_{i3} = (1, \text{KLT6}_i, \text{KGE6}_i, \text{NWINC}_i, \text{MEDU}_i, \text{FEDU}_i, \text{HEDU}_i)',$$

$$w_{i21} = w_{i31} = \text{AGE}_i, \qquad \text{and} \qquad w_{i11} = w_{i22} = w_{i32} = \text{EXPER}_i .$$

Note that this model is a slightly simplified version of the example set out in Section 1 because it does not include on-the-job training, which is an endogenous variable available only for the selected sample. The choice of covariates and instruments in our model is consistent with 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 particular interest because these covariates embody cohort, productivity, and life-cycle effects that are likely to affect wages nonlinearly. Wooldridge (2002, chap. 17) considered 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 6.

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 interferes with a woman's education. The results also show that women who live in families with higher nonwife 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 nonwife 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 estimates of $\mathbf{\Omega}$ provide evidence that education is endogenous: the 95% credibility interval of ω_{21} is mostly in positive territory. In addition, although the errors in the

Table 3. Parameter estimates for nonparametric model of women's wage function model under the priors $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 5 \times \mathbf{I})$, $\boldsymbol{\Omega} \sim \mathcal{IW}(7, 1.2 \times \mathbf{I})$, $g_{2k}|\tau_k^2 \sim \mathcal{N}(0, \tau_k^2/E(\tau_k^2))$, and $\tau_k^2 \sim \mathcal{IG}(6, 0.04)$ for $k = 1, \ldots, 6$. The table also reports 95% credibility intervals and inefficiency factors from 25,000 MCMC iterations.

Parameter	Covariate	Mean	SD	Median	Lower	Upper	Ineff
$\boldsymbol{\beta}_1$	1	0.113	0.360	0.103	−0.558	0.806	43.930
$\boldsymbol{\beta}_2$	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
$\boldsymbol{\beta}_3$	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

log-wage equation are largely uncorrelated with those in the hours equation, sample selection is nonignorable because the correlation between the errors in the education and hours equations is clearly positive.

We now consider the nonparametric functions plotted in Figure 6. These suggest 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 30 and 60 years old, 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 cohort and life-cycle effects, and a strong positive effect of experience on hours, which is consistent with increases in productivity as experience grows.

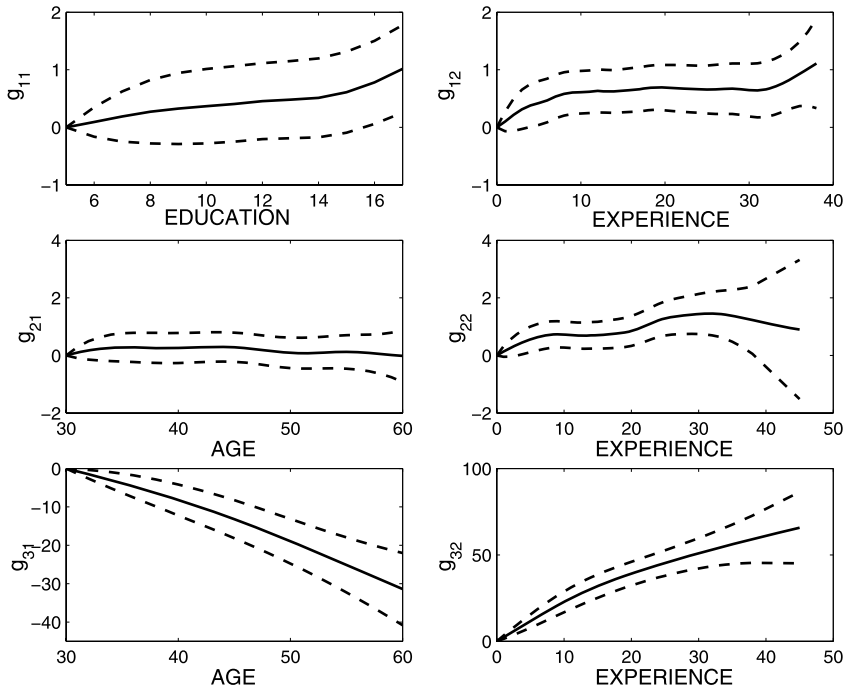

Figure 6. Function estimates in the log-wage application.

Because, with a few exceptions, the nonparametric profiles do not show substantial curvature, we compare this model to several simpler alternatives. The log marginal likelihood of the model with six nonparametric functions is estimated to be $-4{,}236.44$ with a numerical standard error of 0.144, and a model that models only $g_1^1(y_{i2})$ nonparametrically, the other covariates being entered linearly, has an almost identical log marginal likelihood of $-4{,}236.43$ with a numerical standard error of 0.075. Because the data do not provide enough information to distinguish between them, these two models appear equiprobable. These models are also compared to two parametric models—the log marginal likelihood estimate for a linear model is $-4{,}237.405$ 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 $-4{,}244.58$ 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 education and experience, are nonlinear.

7. CONCLUSIONS

This article introduces an efficient approach to analyzing a general class of models in which the problem of sample selection 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 multiequation systems of responses

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

An important aspect of our MCMC algorithm for this class of models is that it does not require simulation of the outcomes that are missing due to the selection mechanism, a feature of the estimation method that enhances computational efficiency as we show in our simulation experiments. Thus, one should not include such missing outcomes in the MCMC simulation when it is possible to proceed otherwise. In our case, even without the inclusion of these missing outcomes, all sampling is from full conditional distributions unless there are constraints on the covariance matrix arising from binary response or binary endogenous variables. In the latter cases, modified MCMC algorithms are available as we point out. 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.

SUPPLEMENTAL MATERIALS

Labor Data: The raw data used in the labor force example. (mroz.raw, text file)

Description of Labor Data: An ascii file describing the contents of file `mroz.raw`. (mroz.des, text file)

Computer Code: This file contains the most important Gauss functions used in the simulations—those for sampling $\mathbf{\Omega}$ and $\{\mathbf{g}_{jk}\}$. (CGJjcgs.s, Gauss code)

ACKNOWLEDGMENT

We thank the Weidenbaum Center on the Economy, Government, and Public Policy for financial assistance.

[Received June 2007. Revised September 2008.]

REFERENCES

Albert, J., and Chib, S. (1993), "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88, 669–679.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (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.

——— (1995), "Marginal Likelihood From the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313–1321.

——— (2001), "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Vol. 5, eds. J. J. Heckman and E. Leamer, Amsterdam: North Holland, pp. 3569–3649.

Chib, S., and Greenberg, E. (1995), "Hierarchical Analysis of SUR Models With Extensions to Correlated Serial Errors and Time Varying Parameter Models," *Journal of Econometrics*, 68, 339–360.

——— (1998), "Analysis of Multivariate Probit Models," *Biometrika*, 85, 347–361.

——— (2007), "Analysis of Additive Instrumental Variable Models," *Journal of Computational and Graphical Statistics*, 16, 86–114.

Chib, S., and Jeliazkov, I. (2006), "Inference in Semiparametric Dynamic Models for Binary Longitudinal Data," *Journal of the American Statistical Association*, 101, 685–700.

Das, M., Newey, W. K., and Vella, F. (2003), "Nonparametric Estimation of Sample Selection Models," *Review of Economic Studies*, 70, 33–58.

Denison, D. G. T., Holmes, C. C., Mallick, B. K., and Smith, A. F. M. (2002), *Bayesian Methods for Nonlinear Classification and Regression*, New York: Wiley.

Denison, D. G. T., Mallick, B. K., and Smith, A. F. M. (1998), "Automatic Bayesian Curve Fitting," *Journal of the Royal Statistical Society*, Ser. B, 60, 333–350.

DiMatteo, I., Genovese, C. R., and Kass, R. E. (2001), "Bayesian Curve-Fitting With Free-Knot Splines," *Biometrika*, 88, 1055–1071.

Fahrmeir, L., and Lang, S. (2001), "Bayesian Inference for Generalized Additive Mixed Models Based on Markov Random Field Priors," *Journal of the Royal Statistical Society*, Ser. C, 50, 201–220.

Fahrmeir, L., and Tutz, G. (1997), *Multivariate Statistical Modelling Based on Generalized Linear Models*, New York: Springer-Verlag.

Gallant, A. R., and Nychka, D. W. (1987), "Semi-Nonparametric Maximum Likelihood Estimation," *Econometrica*, 55, 363–390.

Gersovitz, M., and MacKinnon, J. (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 Horowitz, J. L. (2005), "Nonparametric Methods for Inference in the Presence of Instrumental Variables," *The Annals of Statistics*, 33, 2904–2929.

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, New York: Chapman & Hall.

Hansen, M. H., and Kooperberg, T. (2002), "Spline Adaptation in Extended Linear Models," *Statistical Science*, 17, 2–51.

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.

——— (1979), "Sample Selection Bias as a Specification Error," *Econometrica*, 47, 153–161.

——— (1993), "What Has Been Learned About Labor Supply in the Past Twenty Years?" *The American Economic Review*, 83, 116–122.

Holmes, C. C., Denison, D. G. T., and Mallick, B. K. (2002), "Accounting for Model Uncertainty in Seemingly Unrelated Regressions," *Journal of Computational and Graphical Statistics*, 11, 533–551.

Koop, G., Poirier, D. J., and Tobias, J. (2005), "Bayesian Semiparametric Inference in Multiple Equation Models," *Journal of Applied Econometrics*, 20, 723–747.

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., Wong, W. H., and Kong, A. (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., Rosner, G., Inoue, L., and Dewhirst, M. (2001), "A Bayesian Model for Detecting Acute Change in Nonlinear Profiles," *Journal of the American Statistical Association*, 96, 1215–1222.

Munkin, M. K., and Trivedi, P. K. (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. (1973), "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, 41, 775–788.

——— (1984), "Smoothness Priors and Nonlinear Regression," *Journal of the American Statistical Association*, 79, 609–615.

Shively, T. S., Kohn, R., and Wood, S. (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*, Ser. B, 47, 1–52.

Smith, M., and Kohn, R. (1996), "Nonparametric Regression Using Bayesian Variable Selection," *Journal of Econometrics*, 75, 317–343.

——— (2000), "Nonparametric Seemingly Unrelated Regression," *Journal of Econometrics*, 98, 257–281.

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*, Ser. B, 40, 364–372.

Wasserman, L. (2006), *All of Nonparametric Statistics*, New York: Springer.

Whittaker, E. (1923), "On a New Method of Graduation," *Proceedings of the Edinburgh Mathematical Society*, 41, 63–75.

Wood, S., and Kohn, R. (1998), "A Bayesian Approach to Nonparametric Binary Regression," *Journal of the American Statistical Association*, 93, 203–213.

Wood, S., Kohn, R., Shively, T., and Jiang, W. (2002), "Model Selection in Spline Nonparametric Regression," *Journal of the Royal Statistical Society*, Ser. B, 64, 119–139.

Wooldridge, J. M. (2002), *Econometric Analysis of Cross Section and Panel Data*, Cambridge: MIT Press.