Semiparametric Modeling and Estimation of Instrumental Variable Models

Siddhartha CHIB and Edward GREENBERG

We apply Bayesian methods to a model involving a binary nonrandom treatment intake variable and an instrumental variable in which the functional forms of some of the covariates in both the treatment intake and outcome distributions are unknown. Continuous and binary response variables are considered. Under the assumption that the functional form is additive in the covariates, we develop efficient Markov chain Monte Carlo-based approaches for summarizing the posterior distribution and for comparing various alternative models via marginal likelihoods and Bayes factors. We show in a simulation experiment that the methods are capable of recovering the unknown functions and are sensitive neither to the sample size nor to the degree of confounding as measured by the correlation between the errors in the treatment and response equations. In the binary response case, however, estimation of the average treatment effect requires larger sample sizes, especially when the degree of confounding is high. The methods are applied to an example dealing with the effect on wages of more than 12 years of education.

Key Words: Average treatment effect; Bayes factor; Bayesian inference; Function estimation; Marginal likelihood; Markov chain Monte Carlo; Metropolis-Hastings algorithm.

1. INTRODUCTION

An important issue in statistical inference is the calculation of the effect of a categorical treatment intake on an outcome when the treatment intake is nonrandom. A canonical situation of this type arises when assignment to one of two treatment arms is randomized, but the treatment intake is not the same as the assignment for reasons that are both unobserved and correlated with the outcome. In this situation, the randomized assignment variable is an instrumental variable, a variable that is uncorrelated with unobserved confounders (due to the randomization), and has a direct effect on the observed intake, but has no direct effect on the outcome.

Siddhartha Chib is Harry C. Hartkopf Professor, John M. Olin School of Business, Campus Box 1133, Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63130 (E-mail: *chib@wustl.edu*). Edward Greenberg is Professor Emeritus, Department of Economics, Campus Box 1208, Washington University in St. Louis, 1 Brookings Drive, St. Louis, MO 63130 (E-mail: *edg@artsci.wustl.edu*).

Journal of Computational and Graphical Statistics, Volume 16, Number 1, Pages 1–29
DOI: 10.1198/106186007X180723

Instrumental variable models are the standard approach in the econometrics literature for isolating the effect of the treatment intake in the presence of unmeasured or unobservable confounders. For example, as we discuss in Section 6 (p. 23), suppose that we are interested in the effect on wages of education beyond high school. The problem in finding this effect is that educational attainment is likely to be correlated with such unobserved confounders as an individual's ability or motivation, factors that are also likely to directly affect wages. Progress, however, can be made if an instrumental variable is available. For instance, in our example in Section 6, the instrumental variable measures the proximity to a four-year college. This is arguably a reasonable choice for an instrument because proximity to college is likely to be correlated with the decision to attend college, but not to have a direct effect on wages.

Models organized around instrumental variables are now also used in biostatistics. For example, McClellan, McNeil, and Newhouse (1994) studied whether the application of more intensive treatment of acute myocardial infarction in the elderly reduces mortality. To deal with unobserved confounders that may influence the treatment received and the outcome, McClellan et al. (1994) use the distance to alternative types of hospitals as an instrument. More recently, Greenland (2000) provided an introduction to instrumental variables for epidemiologists.

In this article we consider a class of instrumental variable models that are defined by a joint distribution of the outcome and the intake, conditioned on covariates and the instruments. We depart from the existing literature by relaxing the assumption that the exogenous covariates appear in the treatment and outcome models in parameterized form. We assume only that the functional form is additive on a subset of the covariates. The model is completed by a prior distribution on the various unknowns, including the unknown covariate functions. In contrast to much previous Bayesian work with nonparametric functions, our prior on the functions is proper to ensure that the marginal likelihood of the model, needed to calculate Bayes factors, is well defined.

To outline the setting, suppose that y_i is a continuous outcome on the ith subject ($i \leq n$), x_i is the binary treatment, and z_i is a binary instrument. Then the semiparametric model of interest is given by

$$y_i = \mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta + \varepsilon_i \tag{1.1}$$

$$x_i = I(\mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta + u_i > 0), \tag{1.2}$$

where $\mathbf{v}_{0i} = (v_{01i}, \ldots, v_{0p_0i})'$, $\mathbf{v}_{1i} = (v_{11i}, \ldots, v_{1p_1i})'$, $\mathbf{w}_{0i} = (w_{01i}, \ldots, w_{0q_0i})'$, and $\mathbf{w}_{1i} = (w_{11i}, \ldots, w_{1q_1i})'$ are vectors of exogenous covariates, the covariates $\mathbf{v}_{1i}$ and $\mathbf{w}_{1i}$ are continuous, and $I(A)$ is the indicator function that equals 1 if A is true and 0 otherwise. In keeping with the requirement for a valid instrument, z_i affects the treatment intake, does not directly affect the outcome, and is independent of the errors (ε_i, u_i) given the covariates.

The functions g and f are of the form

$$g(\mathbf{v}_{1i}) = \sum_{k=1}^{p_1} g_k(v_{1ki})$$
$$f(\mathbf{w}_{1i}) = \sum_{k=1}^{q_1} f_k(w_{1ki}),$$

where the component functions g_k and f_k are unknown. We assume additivity of the functions because the fitting of even bivariate functions is a largely unresolved problem at this time. Holmes and Mallick (2003) provided some approaches for bivariate functions, but the theory is in its infancy. Under the additivity assumption, however, it is possible to consider several covariates nonlinearly, and to capture second-order effects, albeit parametrically, by including interactions of the covariates in $\mathbf{v}_{0i}$ or $\mathbf{w}_{0i}$. In addition, we assume that the errors (ε_i, u_i) are jointly Gaussian with covariance matrix $\mathbf{\Omega} = (\omega_{ij})$, where $\omega_{22} = 1$. The parameter ω_{12} in this matrix is a key object of interest and captures the extent of confounding due to unobservable factors that affect both the intake and the outcome. We note that the Gaussian assumption can be relaxed, for example, in the direction of the scale mixture of Gaussian family of distributions, but this extension is not considered here. Finally, $z_i\delta$ in the above model could be replaced with another nonparametric term $h(z_i)$ if the instrument is continuous.

The model for y_i is a shorthand way of writing two marginal distributions, one for y_{i0} and one for y_{i1}, where the former is the outcome when x_i is 0 and the latter is the outcome when x_i is 1. Only one of these values is observed, depending on the value of x_i; the other outcome is the counterfactual. In our model the potential outcome distributions differ only in their means: the mean of y_{i0} is $\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i})$ and that of y_{i1} is $\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + \beta$. The parameter β is the average treatment effect (ATE). It is identified under the exclusion restriction that z_i is not included in the covariates in (1.1), the exogeneity of the covariates and the instrument, and the nondegeneracy of $z_i\delta$.

We also consider another setting in which the outcome is binary. In this model

$$y_i = I(\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta + \varepsilon_i > 0), \tag{1.3}$$

where $\omega_{11} = 1$ and the other variables and functions are as defined above. The ATE conditioned on $(\mathbf{v}_{0i}, \mathbf{v}_{1i})$ is now given by

$$\text{ATE}(\mathbf{v}_{0i}, \mathbf{v}_{1i}) = \Phi(\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + \beta) - \Phi(\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}))$$

from which the unconditional ATE can be obtained by integration over the observed empirical distribution of the covariates $(\mathbf{v}_0, \mathbf{v}_1)$.

In our view, these semiparametric models are a useful extension of the standard IV model with a binary nonrandom treatment variable. Although there is a large frequentist literature on the estimation of IV models and nonparametric models—for example, Efromovich (1999) and Yatchew (2003)—models of the type we discuss have not been considered before. As we show in this article, Bayesian methods in conjunction with the approach of Albert and Chib (1993) are very convenient for dealing with these models. Our work is

thus a contribution to both the Bayesian literature on IV models (e.g., Chib 2003) and the nonparametric Bayesian literature (e.g., Wood and Kohn 1998; Choudhuri, Ghosal, and Roy 2003).

The rest of the article is organized as follows. Section 2 contains the likelihood function, the prior distribution, and details of the MCMC steps for sampling the posterior distribution in the case of continuous response variable, and Section 3 does the same for the binary case. In Section 4 we explain how the MCMC output may be used to compute the marginal likelihoods and Bayes factors for competing model specifications. Section 5 discusses the results of simulation experiments, Section 6 contains an application to real data, and Section 7 has our conclusions.

2. FITTING OF THE MODEL: CONTINUOUS y

2.1 LIKELIHOOD FUNCTION

Suppose that the responses $\mathbf{y} = (y_1, \ldots, y_n)$ and $\mathbf{x} = (x_1, \ldots, x_n)$ of n randomly selected sample subjects have been drawn from the semiparametric IV model (1.1)–(1.2), and let $\mathbf{z} = (z_1, \ldots, z_n)'$ denote the observed values of the instruments, $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \delta, \omega_{11}, \omega_{12})$ the unknown parameter vector, and

$$\begin{aligned} \mathbf{g}_{\cdot i} &= (g_1(v_{11i}), g_2(v_{12i}), \ldots, g_{p_1}(v_{1p_1i}))' : p_1 \times 1 \\ \mathbf{f}_{\cdot i} &= (f_1(w_{11i}), f_2(w_{12i}), \ldots, f_{q_1}(w_{1q_1i}))' : q_1 \times 1 \end{aligned}$$

the vectors of unknown function values for the ith observation. The joint distribution of (y_i, x_i) given $(\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})$ and the covariates, which are omitted in the following expressions to simplify the notation, is

$$p(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}) = f(y_i|\boldsymbol{\alpha}, \beta, \mathbf{g}_{\cdot i}, \omega_{11})[I(x_i = 0)\Pr(x_i = 0|y_i, \boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}) + I(x_i = 1) \times \Pr(x_i = 1|y_i, \boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})]. \quad (2.1)$$

In this expression, the first term on the right-hand side is the marginal distribution of y_{ji}, $j = 0, 1$. We obtain these marginal distributions directly from the specification in (1.1). The terms in square brackets involve the indicator function of the event $x_i = j$ and the conditional probability of each level of the intake, where these are derived from the joint model of the outcome and the intake via some elementary calculations. Following this, we get that the joint distribution in (2.1) is given by

$$f(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}) = \mathcal{N}(y_i|\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta, \omega_{11})p_i^{x_i}(1 - p_i)^{1-x_i}, \quad (2.2)$$

where $\mathcal{N}(\cdot|\cdot, \cdot)$ denotes the normal density function, $p_i = \Phi(m_i/\sqrt{v_i})$, Φ is the cumulative distribution function of the standard normal density, and

$$\begin{aligned} m_i &= \mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta + (\omega_{12}/\omega_{11})(y_i - \mathbf{v}_{0i}'\boldsymbol{\alpha} - g(\mathbf{v}_{1i}) - x_i\beta) \quad (2.3) \\ v_i &= 1 - \omega_{12}^2/\omega_{11}. \quad (2.4) \end{aligned}$$

The joint density of the set of responses $(\mathbf{y}, \mathbf{x})$ given $(\boldsymbol{\theta}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$ is now available as

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n) = \prod_{i=1}^n p(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}), \tag{2.5}$$

where, if the covariate values $\mathbf{v}_{1j\cdot} = (v_{1j1}, \ldots, v_{1jn}) : n \times 1$ and $\mathbf{w}_{1j\cdot} = (w_{1j1}, \ldots, w_{1jn}) : n \times 1$ are distinct and unique for each j, the dimensions of the unknown function vectors $\{\mathbf{g}_{\cdot i}\}_{i=1}^n$ and $\{\mathbf{f}_{\cdot i}\}_{i=1}^n$ are $n \times p_1$ and $n \times q_1$, respectively.

2.2 PRIOR DISTRIBUTION

We complete the model by specifying prior distributions for the unknown functions and for the remaining parameters.

2.2.1 Prior on the Unknown Functions

We place independent, proper, second-order Markov process priors on each of the unknown functions. We first introduce the notation necessary to explain how these priors are specified. Because of the independence assumption, on which we comment below, it suffices to provide the details for a single function, say $\mathbf{g}_j = (g_j(v_{1j1}), \ldots, g_j(v_{1jn}))' : n \times 1$.

To allow for the possibility that the elements of $\mathbf{v}_{1j\cdot}$ are not unique, let $\mathbf{d}_j : m_j \times 1$ denote a vector whose elements are the m_j unique ordered values of $\mathbf{v}_{1j}$,

$$d_{j1} < d_{j2} < \cdots < d_{jm_j},$$

and let the unknown function ordinates at these ordered values be denoted by

$$\tilde{g}_{jk} = g_j(d_{jk}), \quad k = 1, \ldots, m_j.$$

As an identification constraint, we set $\tilde{g}_{j1} = 0$. Such a constraint is required because the level of the function is not identified; intercept terms are always included in $\mathbf{v}_{0i}$ and $\mathbf{w}_{0i}$ to adjust the levels of the functions at the smallest values of their covariates.

Let $\tilde{\mathbf{g}}_j = (\tilde{g}_{j2}, \ldots, \tilde{g}_{jm_j})' : m_j - 1 \times 1$ denote the unrestricted function values. If there are no ties in the covariate values, $m_j = n$; otherwise it is less than n. In most applications there are few ties, so that m_j is not much smaller than n and $\tilde{\mathbf{g}}_j$ is high dimensional. Upon defining $\mathbf{Q}_j : n \times m_j - 1$ as an incidence (or selection) matrix in which row i has a one in column k if $v_{1ji} = d_{jk}$ and zeros in the remainder of the row, we have that $\mathbf{g}_j = \mathbf{Q}_j\tilde{\mathbf{g}}_j$. This linear transformation of $\tilde{\mathbf{g}}_j$ restores the function values to their original, unordered state and sets to zero the components of $\mathbf{g}_j$ that correspond to d_{j1}. The diagonal matrix $\mathbf{Q}_j'\mathbf{Q}_j$ contains the number of observations associated with the corresponding element of $\tilde{\mathbf{g}}_j$. We also define $h_{jk} = d_{jk} - d_{j,k-1}, k = 2, \ldots, m_j$, as the spacing between successive ordered covariate values.

For $k = 2$, we specify $\tilde{g}_{j2} \sim \mathcal{N}(g_{j20}, \tau_j^2 a_j)$, where g_{j20} is a specified hyperparameter and τ_j^2 and a_j are unknown scale parameters. If desired, g_{j20} may be treated as an unknown parameter in a hierarchical structure. The proper prior for $\tilde{g}_{j2}$ leads to a proper prior on the entire set of function ordinates. For $k \geq 3$ we specify a second-order Markov process

prior. To capture the idea that g_j is a smooth function of its argument, we assume that its derivatives do not change rapidly as its argument changes, suggesting that the second derivatives are small (see Shiller 1984, p. 610). Approximating changes in first derivatives by their sample counterparts, we assume that

$$\frac{\tilde{g}_{jk} - \tilde{g}_{j,k-1}}{h_{jk}} - \frac{\tilde{g}_{j,k-1} - \tilde{g}_{j,k-2}}{h_{j,k-1}} = \frac{\xi_{jk}}{\sqrt{h_{jk}}}, \quad \xi_{jk} \sim \mathcal{N}(0, \tau_j^2), \tag{2.6}$$

where $\mathcal{N}(\cdot, \cdot)$ denotes the normal distribution. Everything else being equal, small values of τ_j^2 lead to smoother functions and large values allow g_j to follow the data more closely. Equation (2.6) may be rewritten as

$$\tilde{g}_{jk} = \left(1 + \frac{h_{jk}}{h_{j,k-1}}\right)\tilde{g}_{j,k-1} - \frac{h_{jk}}{h_{j,k-1}}\tilde{g}_{j,k-2} + u_{jk}, \quad u_{jk} \sim \mathcal{N}(0, \tau_j^2 h_{jk}), \tag{2.7}$$

a second-order Markov process prior that flexibly represents smooth functions. Now letting

$$\boldsymbol{\Delta}_j = \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & 0 \\ -\left(1 + \frac{h_{j3}}{h_{j2}}\right) & 1 & 0 & \dots & 0 & \\ \frac{h_{j4}}{h_{j3}} & -\left(1 + \frac{h_{j4}}{h_{j3}}\right) & 1 & 0 & \dots & 0 \\ \ddots & \ddots & \ddots & & & \vdots \\ 0 & \dots & 0 & \frac{h_{jm_j}}{h_{j,m_j-1}} & -\left(1 + \frac{h_{jm_j}}{h_{j,m_j-1}}\right) & 1 \end{pmatrix}$$

be an $m_j - 1 \times m_j - 1$ lower-triangular banded matrix, $\mathbf{g}_{j0} = (g_{j20}, 0, \dots, 0)' : m_j - 1 \times 1$, and $\mathbf{u}_j = (u_{j2}, \dots, u_{jm_j})'$, the vector of independent increments, the form of the joint distribution of $\tilde{\mathbf{g}}_j$ implied by the stochastic process in (2.7) can be written as

$$\boldsymbol{\Delta}_j \tilde{\mathbf{g}}_j = \mathbf{g}_{j0} + \mathbf{u}_j,$$

from which it follows that

$$\tilde{\mathbf{g}}_j | \tau_j^2, a_j \sim \mathcal{N}(\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}, \tau_j^2 \mathbf{K}_j^{-1}), \quad j = 1, \dots, p_1, \tag{2.8}$$

where $\mathbf{K}_j = \boldsymbol{\Delta}_j' \mathbf{H}_j^{-1} \boldsymbol{\Delta}_j$ has full rank and $\mathbf{H}_j = \text{diag}(a_j, h_{j3}, \dots, h_{jm_j})$.

By analogous arguments and notation, we specify the prior on $\tilde{\mathbf{f}}_j$ as

$$\tilde{\mathbf{f}}_j | \tau_{p_1+j}^2, a_{p_1+j} \sim \mathcal{N}\left(\boldsymbol{\Delta}_{p_1+j}^{-1}\mathbf{f}_{j0}, \tau_{p_1+j}^2 \mathbf{K}_{p_1+j}^{-1}\right), \quad j = 1, \dots, q_1. \tag{2.9}$$

Under the assumption of mutual independence, this specification leads to a proper joint prior distribution of $\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_{p_1}, \mathbf{f}_1, \dots, \tilde{\mathbf{f}}_{q_1}$. Without an independence assumption it would be necessary to specify prior covariances between the values of the nonparametric functions at each value of the observed covariates, which requires considerably more a priori information than would be available in practice. Dealing with interdependence between functions, say via nonadditive functions, is not feasible in the current state of knowledge.

As we have seen, the assumption of a second-order Markov prior with Gaussian errors implies that the values of the unknown functions at the observed covariate values have a joint

normal distribution. Many of the priors in the literature share this property. As examples of priors without the enforcement of properness, see Shiller (1973, 1984), Gersovitz and MacKinnon (1978), Besag, Green, Higdon, and Mengersen (1995), Fahrmeir and Tutz (1997, chap. 8), Müller, Rosner, Inoue, and Dewhirst (2001), and Fahrmeir and Lang (2001). Our prior is most similar in design to the last of these and differs principally in our adoption of a proper distribution on g_{j2}. As we have noted, a proper prior permits the use of Bayes factors for model comparison, which may be advantageous in some applications. As always, the choice of prior distribution depends on the state of the researcher's prior knowledge, and the sensitivity of important results to the choice of the prior should be investigated in applications.

In contrast, the first-order prior (e.g., Koop and Poirier 2004) yields functions that are less smooth than those found with our second-order Markov prior, but all priors in the Markov class assume one or more continuous derivatives and can therefore only roughly approximate step functions and other functions with sharply changing levels or slopes. A Gaussian process prior was assumed by Choudhuri et al. (2003). Their approach requires a prior distribution on $\tilde{\mathbf{g}}$, which is specified hierarchically through a polynomial function of the covariate value, and a prior distribution for the covariance matrix. Although the amount of smoothness may be controlled by fixing individual elements in the covariance matrix, they analyze a prior that is based on an exponential function of the distance between observations. Other Bayesian approaches to models with unknown functional forms work with linear combinations of basis functions; for examples in the Bayesian tradition, see Holmes and Mallick (2001, 2003) and Smith and Kohn (2000). In these articles, the functions are also assumed to have considerable smoothness.

It should also be noted that a prior on the nonparametric function values can be specified in terms of the so-called regression splines, where basis functions are specified at knot points and a prior distribution is placed on the coefficients of the basis functions. Recent references to Bayesian work on regression splines are Biller (2000), DiMatteo, Genovese, and Kass (2001), Hansen and Kooperberg (2002), Lang and Brezger (2004), and Baladandayuthapani, Mallick, and Carroll (2005). We plan on comparing the Markov process and regression spline approaches in future research.

In short, like other formulations, our prior implies a Gaussian distribution for the values of the unknown functions. It differs from previous priors by its assumption of properness since we are interested in the calculation of marginal likelihoods and Bayes factors. The assumption of continuous second derivatives is not inconsistent with what is often assumed by researchers in econometric and other applications. If necessary, we can modify our specification for functions with discontinuities by interacting the variable with one or more dummy variables. Another feature of our prior is that it is defined in terms of a small number of parameters and hyperparameters that can be modeled hierarchically, whereas priors based on basis functions or polynomials generally involve more hyperparameters. The simulated examples of Section 5 (p. 15) show that the prior is flexible and in combination with the data is able to approximate three nonlinear functions accurately in both the treatment and response equations.

Finally, we do not claim that all analysis of such problems should be conducted with

our Markov process prior. We have found the prior we use to be easy to formulate and quite capable of capturing various types of functions. In practice it may be of interest to examine the sensitivity of posterior inferences to changes in our prior, and additional experimentation along these lines would be needed for a fuller understanding of the influence of the prior on key results. Our imposition of properness on the prior, however, is an important refinement because it permits us to compare competing models by formal Bayesian methods.

2.2.2 Prior on the Remaining Parameters

We complete our modeling of the prior by specifying a prior distribution for $\boldsymbol{\lambda} = \left(\boldsymbol{\alpha}', \boldsymbol{\gamma}', \delta\right)'$, $\boldsymbol{\tau}^2 = (\tau_1, \ldots, \tau^2_{p_1+q_1})$, $\mathbf{a} = (a_1, \ldots, a_{p_1+q_1})$, β, and the unique elements of $\boldsymbol{\Omega}$. The forms of these priors are chosen to facilitate calculation of the various conditionals needed in the MCMC algorithm.

It is crucial to model the parameters $(\beta, \omega_{11}, \omega_{12})$ jointly because the parameters β and ω_{12} are negatively related—both serve to connect the distribution of the treatment to the response, while ω_{11} and w_{12} are connected through the positive-definiteness constraint on Ω. Upon remparameterizing ω_{11} as $\sigma_{11} = \omega_{11} - \omega^2_{12}$, we assume the prior

$$\pi(\sigma_{11}, \omega_{12}, \beta) = \pi(\sigma_{11})\pi(\omega_{12}, \beta|\sigma_{11}) = \mathcal{IG}\left(\sigma_{11}|\frac{\nu_0}{2}, \frac{\delta_0}{2}\right)\mathcal{N}_2(\omega_{12}, \beta|\mathbf{b}_0, \sigma_{11}\mathbf{B}_0), \tag{2.10}$$

where $\mathcal{IG}(\cdot|\cdot, \cdot)$ is the inverse gamma density and $\mathbf{B}_0$ is a 2×2 matrix with a negative off-diagonal element. Under this reparameterization and prior, the matrix $\boldsymbol{\Omega}$ is positive definite for any value of $\sigma_{11} > 0$ and ω_{12}.

The parameters $(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a})$ are assumed to be mutually independent, with τ^2_j and a_j exchangeable, so that

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}) = \mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\lambda}_0, \mathbf{L}_0)\prod_{j=1}^{p_1+q_1}\mathcal{IG}\left(\tau^2_j|\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right)\mathcal{IG}\left(a_j|\frac{\nu_{000}}{2}, \frac{\delta_{000}}{2}\right), \tag{2.11}$$

where the hyperparameters, the quantities whose last subscript is zero, are known.

2.3 PRIOR-POSTERIOR ANALYSIS

The data consist of the observations $\{y_i, x_i, z_i, \mathbf{v}_{0i}, \mathbf{v}_{1i}, \mathbf{w}_{0i}, \mathbf{w}_{1i}\}$, assumed to be independently drawn across subjects from (1.1)–(1.2). The posterior distribution of the unknown functions and parameters $\boldsymbol{\theta} = (\beta, \sigma_{11}, \omega_{12}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a})$ is given by

$$\pi(\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{p_1}, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_{q_1}, \boldsymbol{\theta}|\mathbf{y}, \mathbf{x})$$
$$\propto \pi(\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{p_1}, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_{q_1}, \boldsymbol{\theta})\prod_{i=1}^{n} p(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}),$$

where the prior is defined above and the density $p(y_i, x_i|\theta, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})$ is that of Equation (2.5). Dependence on the exogenous covariates and the instrumental variable has been suppressed in the notation.

An MCMC algorithm is most conveniently designed by following the latent variable approach of Albert and Chib (1993). We introduce $x_i^* = \mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta + u_i$ and let $x_i = I(x_i^* > 0)$. The posterior density of the latent variables $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$ and the unknown functions and parameters is given by

$$\begin{aligned}\pi(\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{p_1}, \tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_{q_1}, \mathbf{x}^*, \boldsymbol{\theta}|\mathbf{y}, \mathbf{x}) &\propto \prod_{j=1}^{p_1} \mathcal{N}(\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}, \tau_j^2\mathbf{K}_j^{-1}) \\ &\times \prod_{j=1}^{q_1} \mathcal{N}(\boldsymbol{\Delta}_{p_1+j}^{-1}\mathbf{f}_{j0}, \tau_{p_1+j}^2\mathbf{K}_{p_1+j}^{-1}) \mathcal{IG}\left(\sigma_{11}|\frac{\nu_0}{2}, \frac{\delta_0}{2}\right) \\ &\times \mathcal{N}_2(\omega_{12}, \beta|\mathbf{b}_0, \sigma_{11}\mathbf{B}_0)\mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\lambda}_0, \mathbf{L}_0) \prod_{j=1}^{p_1+q_1} \mathcal{IG}\left(\tau_j^2|\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right) \mathcal{IG}\left(a_j|\frac{\nu_{000}}{2}, \frac{\delta_{000}}{2}\right) \\ &\times \prod_{i=1}^{n} p(y_i, x_i^*|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}) I\left\{x_i^* \in C_i\right\}, \end{aligned} \tag{2.12}$$

where $p(y_i, x_i^*|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})$ has the form of a bivariate normal density with location vector

$$\begin{pmatrix} \mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta \\ \mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta \end{pmatrix}$$

and dispersion matrix Ω, and C_i is the interval $(0, \infty)$ if $x_i = 1$ and $(-\infty, 0]$ if $x_i = 0$.

We use MCMC simulation to approximate the posterior distribution in (2.12). The simulation is conducted by sampling blocks of parameters in turn from the so-called full conditional distributions, where each draw is conditioned on the data and the most recent values of the remaining blocks. Sampling sequentially in this way creates a Markov chain whose limiting distribution is the posterior distribution of interest.

In the sequel we define $\boldsymbol{\psi}$ to be the vector of all parameters, latent $\mathbf{x}^*$ and unknown functions and use $\boldsymbol{\psi}\backslash\sigma_{11}$, for example, to denote the vector $\boldsymbol{\psi}$ without the parameter σ_{11}. We also let $\mathbf{V}_0 = (\mathbf{v}_{01}, \ldots, \mathbf{v}_{0n})'$ and $\mathbf{W}_0 = (\mathbf{w}_{01}, \ldots, \mathbf{w}_{0n})'$.

Algorithm 1: **Continuous response IV model**

1. Sample each $\tilde{\mathbf{g}}_j$, the unique values of each unknown function in the response equation, from $\tilde{\mathbf{g}}_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\tilde{\mathbf{g}}_j$.
2. Sample each $\tilde{\mathbf{f}}_j$, the unique values of each unknown function in the treatment equation, from $\tilde{\mathbf{f}}_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\tilde{\mathbf{f}}_j$.
3. Sample each τ_j^2 from $\tau_j^2|\mathbf{y}, \mathbf{x}, \psi\backslash\tau_j^2$.
4. Sample each a_j from $a_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash a_j$.
5. Sample $\mathbf{x}^*$ from $\mathbf{x}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\mathbf{x}^*$.
6. Sample $\boldsymbol{\lambda} = (\boldsymbol{\alpha}, \boldsymbol{\gamma}, \delta)$ from $\boldsymbol{\lambda}|\boldsymbol{\psi}\backslash\boldsymbol{\lambda}$.

7. Sample $(\beta, \sigma_{11}, \omega_{12})$ as a block in two steps: σ_{11} marginalized over (β, ω_{12}) from $\sigma_{11}|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash(\sigma_{11}, \beta, \omega_{12})$, and then (β, ω_{12}) from $\beta, \omega_{12}|\mathbf{x}, \mathbf{y}, \boldsymbol{\psi}\backslash(\beta, \omega_{12})$.
8. Goto 1

The distributions of Steps 1 and 2 are given in the next paragraph, and those in Steps 3–7 are in Appendix A. We also use a variant of this algorithm, called Algorithm 2, in which the first element of $\boldsymbol{\alpha}$—the intercept in the outcome equation—is sampled along with β and ω_{12} in one block.

2.3.1 Sampling the Unknown Functions

These are sampled one function at a time conditioned on the data and the remaining unknowns. Consider, for instance, the sampling of $\tilde{\mathbf{g}}_j$. Conditioned on the data and $\boldsymbol{\psi}$, we express the model in terms of $\tilde{\mathbf{g}}_j$ for all n observations as

$$\tilde{\mathbf{y}}_j|\mathbf{x}, \boldsymbol{\psi} \sim \mathcal{N}_n(\mathbf{Q}_j\tilde{\mathbf{g}}_j, \sigma_{11}\mathbf{I}_n),$$

where

$$\tilde{\mathbf{y}}_j = \mathbf{y} - \mathbf{V}_0\boldsymbol{\alpha} - \sum_{k\neq j}\mathbf{Q}_k\tilde{\mathbf{g}}_k - \mathbf{x}\beta - \omega_{12}(\mathbf{x}^* - \mathbf{W}_0\boldsymbol{\gamma} - \sum_k \mathbf{Q}_k\tilde{\mathbf{f}}_k - \mathbf{z}\delta).$$

It follows from the usual Bayesian calculations that the updated distribution of $\tilde{\mathbf{g}}_j$ is

$$\tilde{\mathbf{g}}_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\tilde{\mathbf{g}}_j \sim \mathcal{N}_{m_j-1}(\hat{\mathbf{g}}_j, \mathbf{G}_j), \tag{2.13}$$

where $\mathbf{G}_j = (\tau_j^{-2}\mathbf{K}_j + \sigma_{11}^{-1}\mathbf{Q}_j'\mathbf{Q}_j)^{-1}$ and $\hat{\mathbf{g}}_j = \mathbf{G}_j(\tau_j^{-2}\mathbf{K}_j\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0} + \sigma_{11}^{-1}\mathbf{Q}_j'\tilde{\mathbf{y}}_j)$. Despite the large dimensions of the matrices in these updates, arising from the fact that m_j can run into the hundreds, Chib and Jeliazkov (2006) showed that all calculations are feasible because of the banded structure of $\boldsymbol{\Delta}_j$ and $\mathbf{K}_j$. These calculations use the fact that the precision matrix $\mathbf{G}_j^{-1}$ is banded because $\mathbf{Q}_j'\mathbf{Q}_j$ is a diagonal matrix. In addition, the quantity $\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}$ is computed once at the outset of the iterations. Also, because $\mathbf{Q}_j'\tilde{\mathbf{y}}_j$ is an $m_j - 1$ vector in which each element is the sum of the $\tilde{\mathbf{y}}_j$ observations associated with the corresponding element of $\tilde{\mathbf{g}}_j$, the quantity $(\tau_j^{-2}\mathbf{K}_j\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0} + \sigma_{11}^{-1}\mathbf{Q}_j'\tilde{\mathbf{y}}_j)$ is readily available. The vector $\hat{\mathbf{g}}_j$ can then be computed by solving the system of equations $\mathbf{G}_j^{-1}\hat{\mathbf{g}}_j = (\tau_j^{-2}\mathbf{K}_j\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0} + \sigma_{11}^{-1}\mathbf{Q}_j'\tilde{\mathbf{y}}_j)$. Finally, a random draw from the given distribution is obtained as $\hat{\mathbf{g}}_j + \mathbf{e}_j$, where $\mathbf{e}_j$ is obtained by solving the equations $\mathbf{L}_j^{*\prime}\mathbf{e}_j = \mathbf{u}_j$, where $\mathbf{L}_j^{*\prime}$ is the cholesky factor of the precision matrix $\mathbf{G}_j^{-1}$ and $\mathbf{u}_j$ is $\mathcal{N}_{m_j-1}(\mathbf{0}, \mathbf{I})$.

The updated distribution of $\tilde{\mathbf{f}}_j$ is obtained analogously, starting with

$$\tilde{\mathbf{x}}_j^* = \mathbf{x}^* - \mathbf{W}_0\boldsymbol{\gamma} - \sum_{k\neq j}\mathbf{Q}_k\tilde{\mathbf{f}}_k - \mathbf{z}\delta - \left(\frac{\omega_{12}}{\omega_{11}}\right)(\mathbf{y} - \mathbf{V}_0\boldsymbol{\alpha} - \sum_k \mathbf{Q}_k\tilde{\mathbf{g}}_k - \mathbf{x}\beta),$$

whose conditional distribution is

$$\tilde{\mathbf{x}}_j^*|\mathbf{y}, \boldsymbol{\psi}\backslash\mathbf{x}^* \sim \mathcal{N}_{m_j-1}(\mathbf{Q}_j\tilde{\mathbf{f}}_j, (1 - \omega_{12}^2/\omega_{11})\mathbf{I}_n).$$

On combining this distribution with the prior of $\tilde{\mathbf{f}}_j$ we obtain

$$\tilde{\mathbf{f}}_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\tilde{\mathbf{f}}_j \sim \mathcal{N}_{m_j-1}(\hat{\mathbf{f}}_j, \mathbf{F}_j), \tag{2.14}$$

where $\mathbf{F}_j = (\tau_j^{-2}\mathbf{K}_j + (1-\omega_{12}^2/\omega_{11})^{-1}\mathbf{Q}_j'\mathbf{Q}_j)^{-1}$ and $\hat{\mathbf{f}}_j = \mathbf{F}_j(\tau_j^{-2}\mathbf{K}_j\Delta_j^{-1}\mathbf{f}_{j0} + (1-\omega_{12}^2/\omega_{11})^{-1}\mathbf{Q}_j'\tilde{\mathbf{x}}_j^*)$.

3. FITTING OF THE MODEL: BINARY y

In this section we briefly explain how the algorithm described in the previous section is modified for binary outcomes. With $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \beta, \boldsymbol{\gamma}, \delta, \omega_{12})$, the joint probability mass function of (y_i, x_i), conditioned on $\boldsymbol{\theta}$, $\mathbf{g}_{\cdot i}$, and $\mathbf{f}_{\cdot i}$ is

$$p(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})$$

$$= \Phi(s_{1i}(\mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta), s_{2i}(\mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta), s_{1i}s_{2i}\omega_{12}), \tag{3.1}$$

where $s_{1i} = 2(y_i - 1)$, $s_{2i} = 2(x_i - 1)$, and $\Phi(a, b, c)$ is the cdf of the bivariate normal distribution evaluated at (a, b) with mean $\mathbf{0}$, unit variances, and covariance c. This model is thus similar to the multivariate probit model considered by Chib and Greenberg (1998), but generalized by the inclusion of the treatment indicator in the outcome specification and by the inclusion of nonparametric functions.

To model the prior information about the parameters $\boldsymbol{\theta}$, let $\boldsymbol{\beta} = (\beta, \omega_{12})$ have the density $\pi(\boldsymbol{\beta}) = \mathcal{N}(\mathbf{b}_0, \mathbf{B}_0)I(-1 < \omega_{12} < 1)$, and independently let $(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a})$ follow the density

$$\pi(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}) = \mathcal{N}(\boldsymbol{\lambda}|\boldsymbol{\lambda}_0, \mathbf{L}_0) \prod_{j=1}^{p_1+q_1} \mathcal{IG}\left(\tau_j^2|\frac{\nu_{00}}{2}, \frac{\delta_{00}}{2}\right)\mathcal{IG}\left(a_j|\frac{\nu_{000}}{2}, \frac{\delta_{000}}{2}\right).$$

Under the same prior distribution on the unknown functions as above and under the assumption that the data are n randomly drawn observations from (3.1), the posterior density of the parameters and unknown function ordinates is

$$\begin{aligned}\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}\,\{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n|\mathbf{y}, \mathbf{x}) &\propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}) \\ &\times \prod_{j=1}^{p_1}\mathcal{N}\left(\tilde{\mathbf{g}}_j|\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}, \tau_j^2\mathbf{K}_j^{-1}\right)\prod_{k=1}^{q_1}\mathcal{N}\left(\tilde{\mathbf{f}}_k|\boldsymbol{\Delta}_k^{-1}\mathbf{f}_{k0}, \tau_{p_1+k}^2\mathbf{K}_k^{-1}\right) \\ &\times \prod_{i=1}^{n} p(y_i, x_i|\boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}).\end{aligned} \tag{3.2}$$

As explained by Albert and Chib (1993) and Chib and Greenberg (1998), simulation from this posterior density is facilitated by the introduction of latent variables y_i^* to represent the discrete outcomes y_i, where $y_i^* = \mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta + \varepsilon_i$, with $\varepsilon_i \sim \mathcal{N}(0, 1)$ and $y_i = I(y^* > 0)$. In terms of the latent variables $\mathbf{y}^* = (y_1^*, \ldots y_n^*)$ and $\mathbf{x}^*$ the posterior

density is

$$\pi(\boldsymbol{\beta}, \boldsymbol{\lambda}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n, \mathbf{y}^*, \mathbf{x}^* | \mathbf{y}, \mathbf{x}) \propto \pi(\boldsymbol{\beta})\pi(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}) \times \prod_{j=1}^{p_1} \mathcal{N}\left(\tilde{\mathbf{g}}_j | \boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}, \tau_j^2 \mathbf{K}_j^{-1}\right) \prod_{k=1}^{q_1} \mathcal{N}\left(\tilde{\mathbf{f}}_k | \boldsymbol{\Delta}_k^{-1}\mathbf{f}_{k0}, \tau_{p_1+k}^2 \mathbf{K}_k^{-1}\right) \times \prod_{i=1}^{n} p(y_i^*, x_i^* | \boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i}) I\left\{\left(y_i^*, x_i^*\right) \in C_i\right\}, \tag{3.3}$$

where $p(y_i^*, x_i^* | \boldsymbol{\theta}, \mathbf{g}_{\cdot i}, \mathbf{f}_{\cdot i})$ has the form of a bivariate normal density with location vector

$$\begin{pmatrix} \mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta \\ \mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta \end{pmatrix}$$

and dispersion matrix Ω with units on the diagonal and ω_{12} on the off-diagonal, and C_i is a subset of R^2 depending on the values (y_i, x_i). For example, if $y_i = 1$ and $x_i = 1$, then $C_i = (0, \infty) \times (0, \infty)$.

Inspection of the posterior density in (3.3) reveals that, conditioned on the latent variables $(\mathbf{y}^*, \mathbf{x}^*)$, the posterior has the same form as that of the continuous outcome model. Accordingly, the same conditional densities arise in the MCMC updates of $\boldsymbol{\lambda}$, $\boldsymbol{\tau}^2$, $\mathbf{a}$, $\{\mathbf{g}_{\cdot i}\}_{i=1}^n$, $\{\mathbf{f}_{\cdot i}\}_{i=1}^n$ by replacing $\mathbf{y}$ everywhere with $\mathbf{y}^*$ and setting ω_{11} to one. A new step is the update of the parameters $\boldsymbol{\beta} = (\beta, \omega_{12})$ by the M-H algorithm (Chib and Greenberg 1995). Rather than update $\boldsymbol{\beta}$ from the posterior density augmented with the latent data, it is possible to update it directly from the posterior density in (3.2), which tends to improve the mixing of the resulting Markov chain. The details are as follows: Let $q(\boldsymbol{\beta}^\dagger) = T_2(\bar{\boldsymbol{\beta}}, \mathbf{V}, 10)$ denote a bivariate-t proposal density with location vector $\bar{\boldsymbol{\beta}}$, obtained by maximizing the log of the posterior density in (3.2), dispersion matrix $\mathbf{V}$ equal to the negative inverse Hessian matrix of the log posterior density evaluated at $\bar{\boldsymbol{\beta}}$, and ten degrees of freedom. Now let $\boldsymbol{\beta}^\dagger$ be a proposal value generated from $q(\boldsymbol{\beta}^\dagger)$ and let $\boldsymbol{\beta}^c$ be the current value of $\boldsymbol{\beta}$. The M-H update is implemented by moving to the proposed value $\boldsymbol{\beta}^\dagger$ with probability $\alpha(\boldsymbol{\beta}^c, \boldsymbol{\beta}^\dagger | \mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$ and staying at the current value with probability $1 - \alpha(\boldsymbol{\beta}^c, \boldsymbol{\beta}^\dagger | \mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$, where

$$\alpha(\boldsymbol{\beta}^c, \boldsymbol{\beta}^\dagger | \mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n) = \min\left\{1, \frac{\pi(\boldsymbol{\beta}^\dagger, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n | \mathbf{y}, \mathbf{x})}{\pi(\boldsymbol{\beta}^c, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n | \mathbf{y}, \mathbf{x})} \frac{q\left(\boldsymbol{\beta}^c\right)}{q\left(\boldsymbol{\beta}^\dagger\right)}\right\}. \tag{3.4}$$

Finally, one iteration of the MCMC algorithm is completed by sampling the latent variables: for the ith observation, sampling (y_i^*, x_i^*) from univariate truncated normal distributions conditioned on all the unknowns of the augmented model. Specifically, with

$$\mu_{1i} = \mathbf{v}_{0i}'\boldsymbol{\alpha} + g(\mathbf{v}_{1i}) + x_i\beta + \omega_{12}(x_i^* - \mathbf{w}_{0i}'\boldsymbol{\gamma} - f(\mathbf{w}_{1i}) - z_i\delta),$$

y_i^* is sampled from $\mathcal{N}(\mu_{1i}, 1 - \omega_{12}^2)$ truncated from above at 0 when $y_i = 0$ and from below at 0 when $y_i = 1$. And with

$$\mu_{2i} = \mathbf{w}_{0i}'\boldsymbol{\gamma} + f(\mathbf{w}_{1i}) + z_i\delta + \omega_{12}(y_i^* - \mathbf{v}_{0i}'\boldsymbol{\alpha} - g(\mathbf{v}_{1i}) - x_i\beta),$$

x_i^* is sampled from $\mathcal{N}(\mu_{2i}, 1 - \omega_{12}^2)$ truncated to the interval $(0, \infty)$ when $x_i = 1$, and truncated to the interval $(-\infty, 0]$ when $x_i = 0$.

4. MODEL COMPARISONS

We now present an approach for comparing our semiparametric model with various alternative models, for example, models with no unobserved confounding, that is, $\omega_{12} = 0$, models distinguished by the presence or absence of specific covariates in the outcome and treatment assignment equations, and models in which some or all of the nonparametric functions appear in an explicit functional form, say linearly or as a low-order polynomial. An important advantage of the Bayesian perspective is that it provides a systematic strategy for comparing these alternatives through the computation of marginal likelihoods and Bayes factors (Jeffreys 1961, chap. 5). As has been demonstrated in numerous articles, the method of Chib (1995) makes it possible to find the marginal likelihood with a modest amount of effort. This method is not restricted to the parametric setting and can be used to compare distribution-free Bayesian models (Basu and Chib 2003). We show that it can also be adapted for semiparametric models with unknown functional form components.

By definition, the marginal likelihood is given by

$$m(\mathbf{y}, \mathbf{x}) = \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}, \tag{4.1}$$

where

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta})\pi(\tilde{\mathbf{g}}|\boldsymbol{\theta})\pi(\tilde{\mathbf{f}}|\boldsymbol{\theta})\, d\tilde{\mathbf{g}}\, d\tilde{\mathbf{f}} \tag{4.2}$$

is the density of the responses marginalized over the unknown functions, $\tilde{\mathbf{g}} = (\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_{p_1})$, and $\tilde{\mathbf{f}} = (\tilde{\mathbf{f}}_1, \ldots, \tilde{\mathbf{f}}_{q_1})$. Although direct calculation of this quantity is not possible, it can be obtained indirectly: Define $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \sigma_{11}, \omega_{12}, \beta, \boldsymbol{\tau}^2, \mathbf{a})$, and let $\pi(\boldsymbol{\theta})$ be its prior density given by the product of the densities (2.10) and (2.11). Also let

$$\pi(\tilde{\mathbf{g}}|\boldsymbol{\theta}) = \prod_{j=1}^{p_1} \mathcal{N}_{m_j - 1}\left(\tilde{\mathbf{g}}_j|\boldsymbol{\Delta}_j^{-1}\mathbf{g}_{j0}, \tau_j^2\mathbf{K}_j^{-1}\right),$$

and

$$\pi(\tilde{\mathbf{f}}|\boldsymbol{\theta}) = \prod_{j=1}^{q_1} \mathcal{N}_{m_j - 1}\left(\tilde{\mathbf{f}}_j|\boldsymbol{\Delta}_j^{-1}\mathbf{f}_{j0}, \tau_{p_1+j}^2\mathbf{K}_j^{-1}\right)$$

denote the conditional prior densities of the unknown functions. Let $\boldsymbol{\theta}^*$ denote the posterior mean of $\boldsymbol{\theta}$ estimated from the MCMC output. Following Chib (1995), we then have that

$$m(\mathbf{y}, \mathbf{x}) = \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*)\pi(\boldsymbol{\theta}^*)}{\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x})},$$

where $\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x})$ is the posterior density, $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*)$ is the likelihood function, and $\pi(\boldsymbol{\theta}^*)$ is the prior density, each evaluated at $\boldsymbol{\theta}^*$. An estimate of the marginal likelihood is found by separately estimating the ordinates $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*)$ and $\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x})$.

Estimation of $\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x})$: Consider first the case of the model with a continuous outcome. Express the posterior ordinate as

$$\pi(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}) = \pi(\boldsymbol{\tau}^{2*}|\mathbf{y}, \mathbf{x})\pi(\sigma_{11}^*, \omega_{12}^*, \beta^*|\mathbf{y}, \mathbf{x})\pi(\mathbf{a}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\tau}^{2*})\pi(\boldsymbol{\lambda}^*|\mathbf{y}, \mathbf{x}, \sigma_{11}^*, \omega_{12}^*, \beta^*), \tag{4.3}$$

in which the first two terms can be estimated from the output of the full MCMC run by averaging the appropriate densities over the simulated draws. Specifically, letting l index the MCMC iteration, we have the simulation-consistent estimate

$$\hat{\pi}(\boldsymbol{\tau}^{2*}|\mathbf{y}, \mathbf{x}) = M^{-1}\sum_{l=1}^{M}\prod_{j=1}^{p_1+q_1}\mathcal{IG}\left(\tau_j^{2*}|\frac{\nu_{00}+m_j-1}{2}, \frac{\delta_{00}+b_j^l}{2}\right),$$

where M is the number of retained iterations and b_j is defined in the Appendix under (A.1). Likewise,

$$\hat{\pi}(\sigma_{11}^*, \omega_{12}^*, \beta^*|\mathbf{y}, \mathbf{x}) = M^{-1}\sum_{l=1}^{M}\mathcal{IG}\left(\sigma_{11}^*|\frac{\nu_0+n}{2}, \frac{\delta_0+d^l}{2}\right)\mathcal{N}(\boldsymbol{\beta}^*|\mathbf{b}_1^l, \sigma_{11}^*\mathbf{B}_1^l),$$

where d, $\mathbf{b}_1$, and $\mathbf{B}_1$ are defined in the text surrounding (A.4) and (A.5). The quantities b_j^l, d^l, $\mathbf{b}_1^l$, and $\mathbf{B}_1^l$ are computed in the course of the MCMC sampling and stored for use in this step.

The two remaining ordinates are now estimated by the reduced run method of Chib (1995): Fix $\boldsymbol{\tau}^2$, β, σ_{11}, and ω_{12} at their starred values, and continue the MCMC simulation with the quantities $\boldsymbol{\psi}\backslash(\boldsymbol{\tau}^2, \beta, \sigma_{11}, \omega_{12})$. The draws from this run are used to average the inverse gamma density of $\mathbf{a}$ in (A.2) and the normal density of $\boldsymbol{\lambda}$ at $\boldsymbol{\lambda}^*$ in (A.3) for continuous y and the analogous expression for binary y. Our estimate of the posterior ordinate is obtained by inserting these estimates into (4.3).

Now in the case of a binary outcome, the second ordinate in (4.3) is $\pi(\omega_{12}^*, \beta^*|\mathbf{y}, \mathbf{x})$ which cannot be estimated by the foregoing method because $\boldsymbol{\beta} = (\omega_{12}^*, \beta^*)$ is sampled from a conditional density (marginalized over the latent variables) that is not in closed form. Instead we use the method of Chib and Jeliazkov (2001). In particular, from one of their results we find that

$$\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}) = \frac{E_1\alpha(\boldsymbol{\beta}, \boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)q(\boldsymbol{\beta}^*)}{E_2\alpha(\boldsymbol{\beta}^*, \boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)},$$

where the expectation E_1 is with respect to the joint posterior of $(\boldsymbol{\beta}, \boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$ and the expectation E_2 is with respect to the posterior distribution of $(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$ given $\boldsymbol{\beta}^*$ times that of the proposal distribution of $\boldsymbol{\beta}$ given $(\boldsymbol{\lambda}, \boldsymbol{\tau}^2, \mathbf{a}, \{\mathbf{g}_{\cdot i}\}_{i=1}^n, \{\mathbf{f}_{\cdot i}\}_{i=1}^n)$. The numerator expectation is estimated from the output of the full run and the denominator expectation from the output of a reduced run given $\boldsymbol{\beta}^*$ in which an appended step involving the sampling of $\boldsymbol{\beta}$ from the proposal distribution is included. The remaining ordinates are estimated as in the continuous outcome case.

Estimation of $p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*)$*:* The likelihood ordinate is also estimated by the approach of Chib (1995). The starting point is the identity

$$p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*) = \frac{p(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*, \tilde{\mathbf{g}}^*, \tilde{\mathbf{f}}^*)\pi(\tilde{\mathbf{g}}^*|\boldsymbol{\theta}^*)\pi(\tilde{\mathbf{f}}^*|\boldsymbol{\theta}^*)}{\pi(\tilde{\mathbf{g}}^*, \tilde{\mathbf{f}}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*)},$$

where $\tilde{\mathbf{g}}^*$ and $\tilde{\mathbf{f}}^*$ denote, for example, the posterior means of $\tilde{\mathbf{g}}$ and $\tilde{\mathbf{f}}$, respectively. The quantities in the numerator are available from (2.5), (2.8), and (2.9). What remains is the computation of the ordinate in the denominator. Proceeding as with $\boldsymbol{\theta}$, employ a marginal/conditional decomposition, writing $\pi(\tilde{\mathbf{g}}^*, \tilde{\mathbf{f}}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*)$ as

$$\pi(\tilde{\mathbf{g}}_1^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*)\pi\left(\tilde{\mathbf{g}}_2^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*, \tilde{\mathbf{g}}_1^*\right) \ldots \pi\left(\tilde{\mathbf{f}}_{q_1}^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*,\ \tilde{\mathbf{g}}^*, \tilde{\mathbf{f}}^* \backslash \tilde{\mathbf{f}}_{q_1}^*\right).$$

Each term is estimated from the output of an appropriate reduced MCMC run, keeping $\boldsymbol{\theta}$ fixed at $\boldsymbol{\theta}^*$ in each run. To estimate the first ordinate, for example, continue the MCMC simulations with the quantities $\tilde{\mathbf{g}}$, $\tilde{\mathbf{f}}$, and $\mathbf{x}^*$ and use the draws from this run to average the density given in (2.13) for continuous y, which yields the estimate

$$\pi(\tilde{\mathbf{g}}_1^*|\mathbf{y}, \mathbf{x}, \boldsymbol{\theta}^*) = M^{-1} \sum_{l=1}^{M} \mathcal{N}_{m_1-1}(\tilde{\mathbf{g}}_1^*|\hat{\mathbf{g}}_1, \mathbf{G}_1),$$

where $\hat{\mathbf{g}}_1$ and $\mathbf{G}_1$ are defined below (2.13). Analogous expressions described in Section 3 are employed for binary y. The remaining ordinates are estimated in the same way from the output of subsequent reduced MCMC runs, where the functions in the conditioning set are set to their starred values. This completes the estimation of the likelihood function.

An estimate of the marginal likelihood on the log scale is now given by

$$\log \hat{m}(\mathbf{y}, \mathbf{x}) = \log \hat{p}(\mathbf{y}, \mathbf{x}|\boldsymbol{\theta}^*) + \log \pi(\boldsymbol{\theta}^*) - \log \hat{\pi}(\boldsymbol{\theta}^*|\mathbf{y}, \mathbf{x}),$$

and the Bayes factor for any two models by the ratio of their respective marginal likelihoods.

5. SIMULATION EXPERIMENTS

We next present the results of a simulation experiment designed to determine whether our MCMC procedure is capable of recovering nonlinear functions for sample sizes typically found in cross-section datasets, whether the number of nonlinear functions included affects results, and whether results are sensitive to the degree of confounding measured by ω_{12}. In detail, we specified the following functions,

$$g_1(v) = 1.5\,(\sin(\pi v))^2,$$

$$g_2(v) = \left(\sin(2\pi v^3)\right)^3,$$

$$g_3(v) = 6\left(1 - \cos((\pi v/4)^2)\right),$$

$$f_1(w) = 6w^3\left(1 - w^3\right),$$

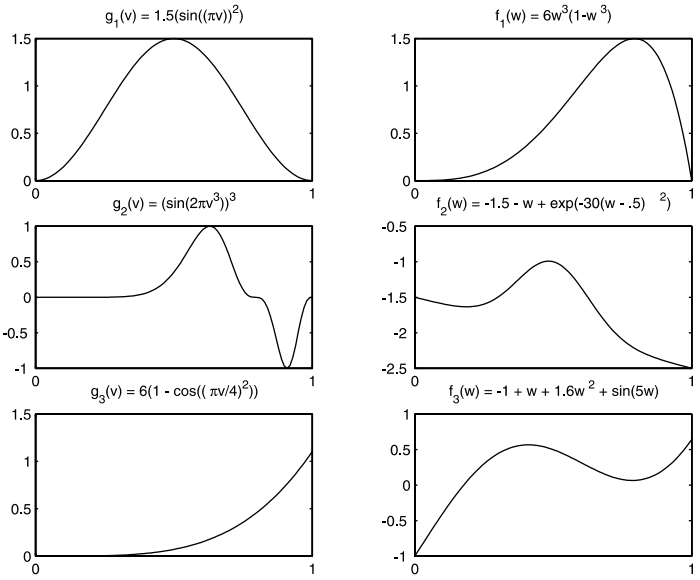

Figure 1. Functions used in simulation experiment.

$$f_2(w) = -1.5 - w + \exp\left(-30(w - .5)^2\right),$$

and

$$f_3(w) = -1 + w + 1.6w^2 + \sin(5w).$$

These functions, graphed in Figure 1, display a wide variety of nonlinear shapes. We set $\omega_{11} = 1$ and considered three different values of ω_{12}: 0.1, 0.5, and 0.9. Sample sizes are set at 1,500 and 2,500, and 20 replications of each design are generated. The covariate vectors are $\mathbf{w}_{0i} = \mathbf{v}_{0i} = (1, w_{02i})'$, where w_{02i} is uniformly distributed on (0, 1). The $w_{11i} = v_{11i}$, $w_{12i} = v_{12i}$, and $w_{13i} = v_{13i}$ covariates are randomly sampled with equal probabilities, respectively, from

$$(0 : 0.05 : 0.1) \cup (0.15 : 0.02 : 0.3) \cup (0.32 : 0.01 : 0.85) \cup (0.86 : 0.03 : 1),$$

$$(0 : 0.025 : 0.25) \cup (0.252 : 0.0125 : 0.5) \cup (0.501 : 0.01 : 0.69) \cup (0.70 : 0.03 : 1),$$

and

$$(0 : 0.03 : 0.2) \cup (0.23 : 0.02 : 0.5) \cup (0.52 : 0.01 : 0.9) \cup (0.91 : 0.02 : 1),$$

where $(a : b : c)$ denotes the sequence $a + b, a + 2b, \ldots, c$. Data are generated from model (1.1)–(1.2) with parameter values listed in Table 1, and z_i is generated from a Bernoulli distribution with parameter 0.6. To study the sensitivity of the procedure to the number of included nonlinear functions, we specify $g(\mathbf{v}_{1i})$ and $f(\mathbf{w}_{1i})$ in three different ways, with one, two, or three functions:

$$g(\mathbf{v}_{1i}) = g_1(v_{11i}), \quad f(\mathbf{w}_{1i}) = f_1(w_{11i});$$

$$g(\mathbf{v}_{1i}) = g_1(v_{11i}) + g_2(v_{12i}), \quad f(\mathbf{w}_{1i}) = f_1(w_{11i}) + f_2(w_{12i});$$

$$g(\mathbf{v}_{1i}) = g_1(v_{11i}) + g_2(v_{12i}) + g_3(v_{13i}), \quad f(\mathbf{w}_{1i}) = f_1(w_{11i}) + f_2(w_{12i}) + f_3(w_{13i}).$$

Table 1. True Values and Prior Moments: Continuous y

		Prior	
Parameter	True value	Mean	Std.Dev.
α_1	2.000	0.000	10.000
α_2	1.000	0.000	3.162
β	1.000	0.500	10.000
γ_1	0.000	0.000	3.162
γ_2	0.500	0.000	3.162
δ	0.500	0.000	3.162
σ_{11}	$1-\omega_{12}^2$	0.500	1.500
ω_{12}	Various	0.150	0.939
τ_j^2 (all j)	n.a.	0.050	0.100
a_j (all j)	n.a.	1.000	1.000

5.1 ESTIMATION RESULTS FOR CONTINUOUS y

Table 1 contains the true values of the various parameters and the hyperparameter values for the prior distributions. For prior distributions we specified neutral means and large standard deviations. The prior distributions should play a small role because we consider only large sample sizes. We discard the first 500 samples from the MCMC algorithm and base the results on the subsequent 10,000 iterations. Results are reported for Algorithm 2 (see p. 9).

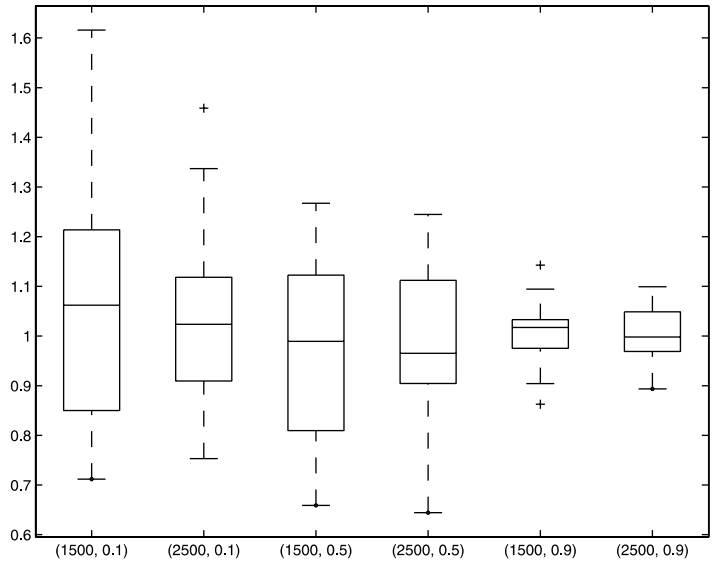

Figure 2. Boxplots of β (true value is 1) for indicated combinations of (n, ω_{12}), three included functions, continuous y.

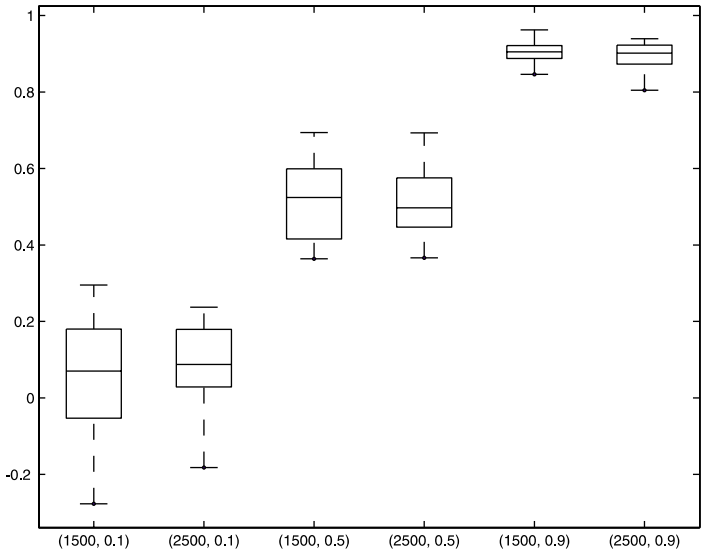

Figure 3. Boxplots of ω_{12} for indicated combinations of (n, ω_{12}), three included functions, continuous y.

Complete tables and graphs of the results are available from the authors. We concentrate on results for the parameters of greatest interest, β and ω_{12}; they are summarized in Figures 2 and 3 for the model that contains three functions. These parameters are well determined, with little sensitivity to sample size, number of functions, or the degree of confounding as measured by ω_{12}. The proportion of replications for which the three parameters are in the 95% credibility interval averages about 95%. As expected, standard deviations and lengths of credibility intervals fall as the sample size increases from 1,500 to 2,500. Less expected is that these also fall as ω_{12} increases from 0.1 to 0.9. That is, β and ω_{12} are better determined as the extent of confounding increases. As for the mixing of the MCMC output, the serial correlations generally decline to zero by lag 20 except for that of ω_{12} whose autocorrelations tend to decline to zero a bit less quickly.

We illustrate results of estimating the nonlinear functions with the challenging case of the model with three functions, $n = 1{,}500$, and $\omega_{12} = 0.9$; Figure 4 graphs g_2 and Figure 5 graphs f_2 for each of the 20 samples. We see that the estimated functions follow the curvature of the true functions for these cases, which are representative of the complete results. This is especially remarkable for f_2, where the only information available to the program is whether x_i is zero or one. (Because of its nonzero intercept, this function has been shifted up to facilitate comparison with the true function.) From this sketchy information the procedure is able to reproduce the true function with considerable accuracy. The complete results reveal no obvious degradation in performance as the number of nonlinear functions is increased from one in each equation to three, nor is there any apparent sensitivity to ω_{12} or the sample size. It is important to recognize how the design affects the results: the relatively poor performance of estimating g_2 in the region between 0.8 and 1.0 is partly due to the scarcity of sample points in that interval—w_{21} in the range from 0.7 to 1.0 takes on

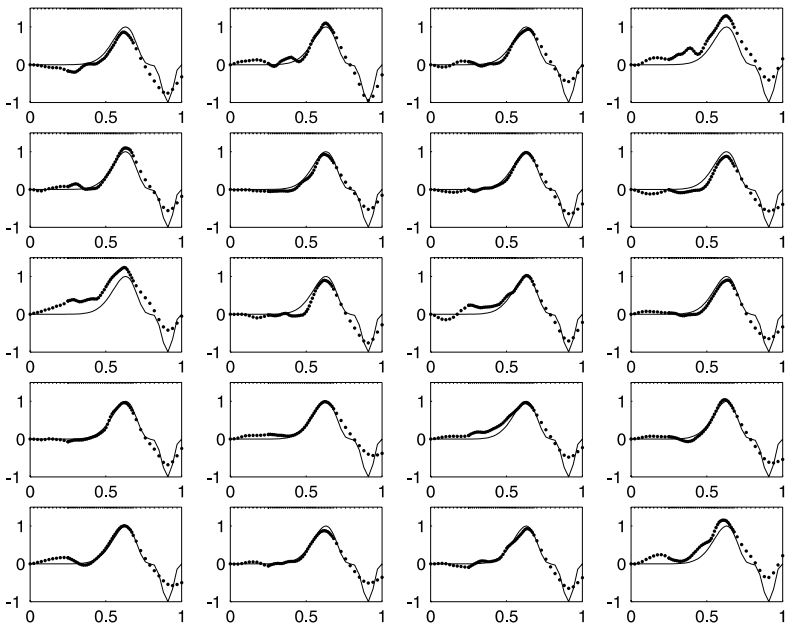

Figure 4. Individual sample results for g_2, continuous y, $n = 1{,}500$, $\omega_{12} = 0.9$ (true functions in solid lines; estimated functions in dotted lines).

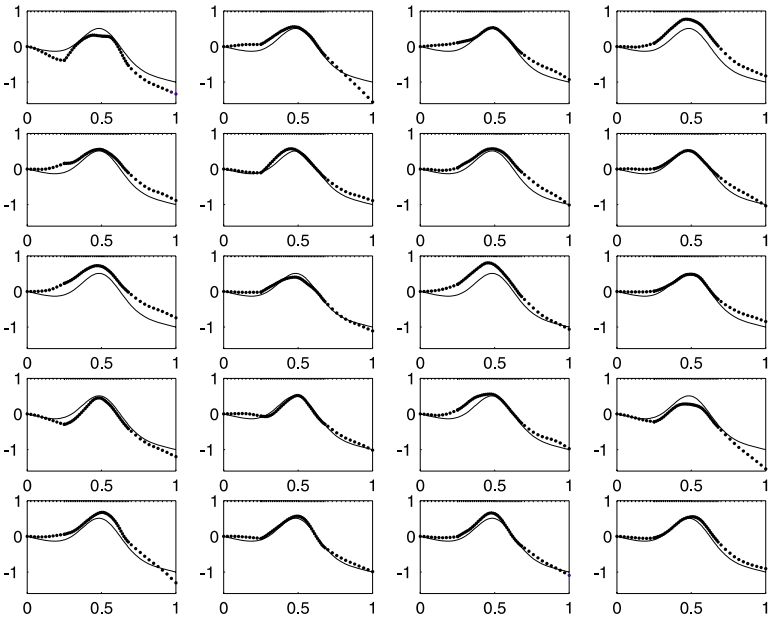

Figure 5. Individual sample results for f_2, continuous y, $n = 1{,}500$, $\omega_{12} = 0.9$ (true functions in solid lines; estimated functions in dotted lines).

Table 2. True Values and Prior Moments: Binary *y*

		Prior	
Parameter	*True value*	*Mean*	*Std. dev.*
α_1	−0.500	−2.000	10.000
α_2	0.500	0.000	3.162
β	0.500	0.000	10.000
γ_1	0.500	0.000	3.162
γ_2	0.500	0.000	3.162
δ	0.500	0.000	3.162
ω_{12}	Various	0.150	1.000
τ_j^2 (all *j*)	n.a.	0.500	0.100
a_j (all *j*)	n.a.	1.000	1.000

values with an increment of 0.03, so that it has only 11 different values in a region where the function changes direction from decreasing to increasing. The scarcity of points is seen by the upper tick marks in the Figures, which indicate covariate values included in the sample.

5.2 ESTIMATION RESULTS FOR BINARY *y*

The binary y model for the simulation experiment is the same as specified above for the continuous case, except that true parameter values and prior parameters are specified in Table 2, $\omega_{11} = 1$, and w_{01i} is uniformly distributed on $(-1, 1)$.

Summary results for β and ω_{12} are in Figures 6 and 7. In contrast to the case of continuous

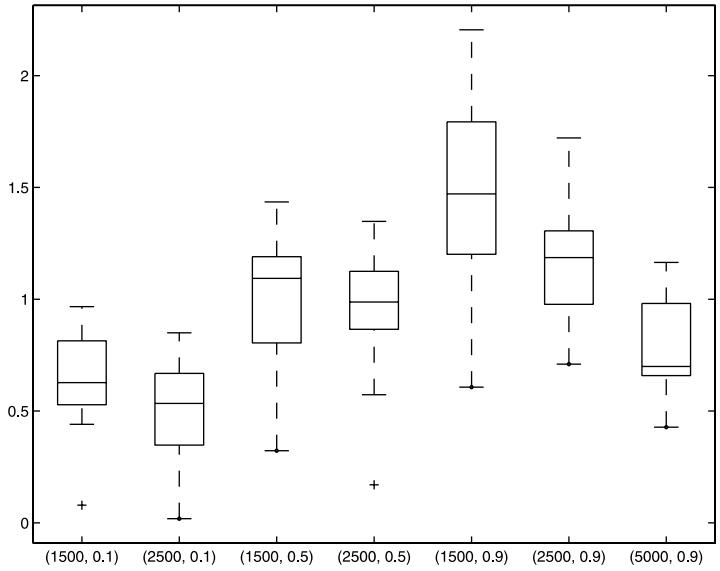

Figure 6. Boxplots of β (true value is 0.5) for indicated combinations of (n, ω_{12}), three included functions, binary y.

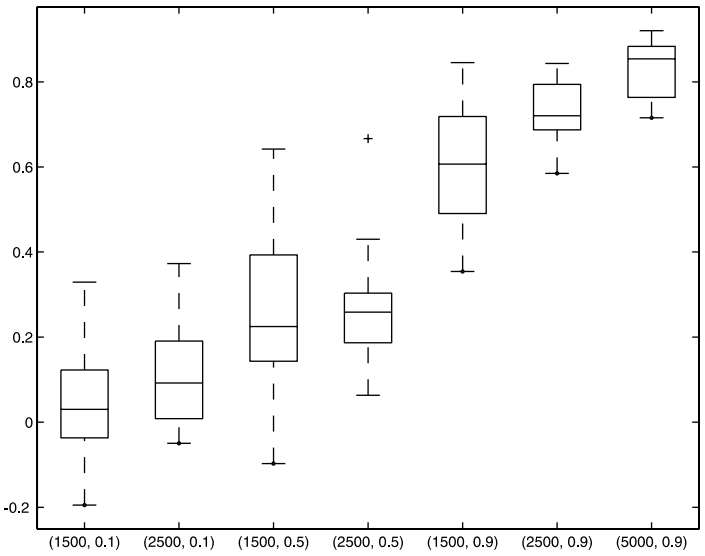

Figure 7. Boxplots of ω_{12} for indicated combinations of (n, ω_{12}), three included functions, binary y.

y, the estimation of β and ω_{12} is quite dependent on values of the latter. The performance of the estimators, as indicated by the closeness of the posterior mean to its true value and by the proportion of the replications for which they are between the lower and upper bounds, degrades as ω_{12} increases. The procedure tends to overestimate β and underestimate ω_{12} for the two larger values of ω_{12}. An increase in sample size moves both parameters in the right direction, but the results are not reliable even for the larger sample size. Much larger sample sizes may be needed to offset the paucity of information available in this case, where both y and x are observed only as binary variables. To investigate this possibility we ran an additional simulation with $n = 5{,}000$ for the most difficult case of three functions in each equation and $\omega_{12} = 0.9$. The posterior median of β falls from 1.4706 to 1.1861 to 0.6992 as the sample size rises from 1,500 to 2,500 to 5,000; it appears to be approaching its true value of 0.5, but rather slowly. Similarly, the median of ω_{12} rises from 0.6065 to 0.7204 to 0.8539, approaching its true value of 0.9. Figures 8 and 9, respectively, graph g_2 and f_2 for the 20 individual samples. Although results in some samples are better than in others, most of the estimates follow the actual functions reasonably well. In viewing these graphs it is important to recall that data for both y_i and x_i are binary, so that the estimated function values are based on the y_i^* and x_i^* generated from the MCMC simulations. As before, there is little sensitivity to ω_{12} or the sample size.

5.3 SIMULATION RESULTS FOR MARGINAL LIKELIHOODS

To study the ability of the marginal likelihood calculation to pick out the true model, we generated 20 datasets from a model that uses the specification of Section 5.1 for continuous y and Section 5.2 for binary y for the parameter values, but includes only g_1 and g_2 in the y or y^* equation and only f_1 and f_2 in the x^* equation. We then fit three models: one with

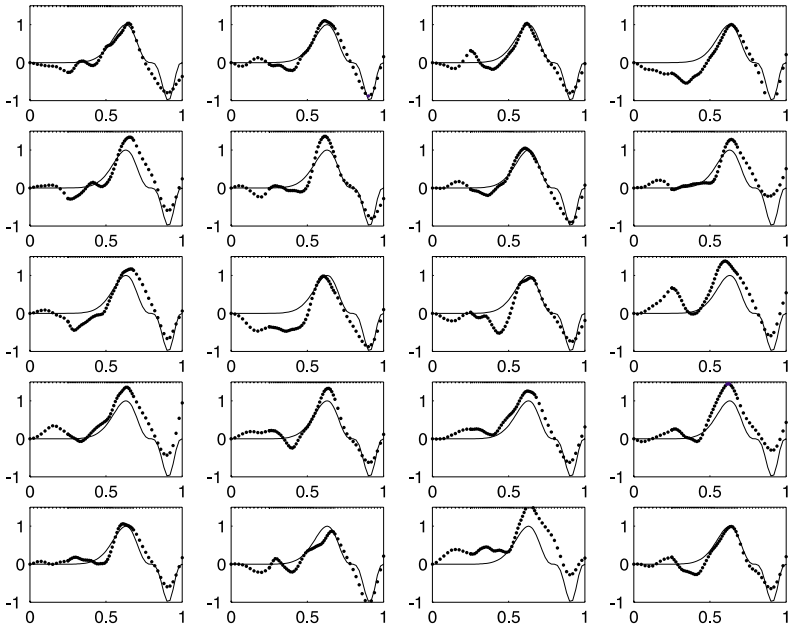

Figure 8. Individual sample results for g_2, binary y, $n = 1{,}500$, $\omega_{12} = 0.9$ (true functions in solid lines; estimated functions in dotted lines).

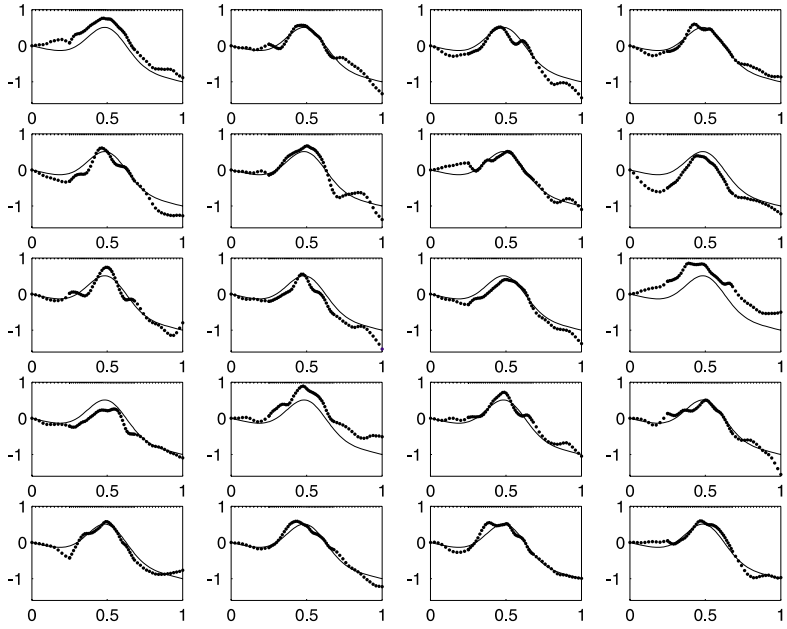

Figure 9. Individual sample results for f_2, binary y, $n = 1500$, $\omega_{12} = 0.9$ (true functions in solid lines; estimated functions in dotted lines).

the correct specification, a second that omits g_2 (resp. f_2) from the y (x^*) equation, and a third that adds g_3 (resp. f_3) to the y (x^*) equation. We set $n = 1{,}500$ on the assumption that obtaining good results with the smaller sample size would be a strong indication that good results would be obtained with the larger, and we set $\omega_{12} = 0.9$, which we consider to be the most challenging case.

Our results for the continuous case show that the correct model is decisively chosen in each of the 20 replications. The Bayes factor for the correct model over the model with three functions in each equation is at least 10^5, and it is at least 10^{85} compared to the model with one function in each equation. For the binary response case, from 20 replications of the experiment we again find that the Bayes factors strongly favor the true model in each replication. The Bayes factors are at least 10^{25} for the true model against the model with one function and at least 10^8 against the model with three functions.

6. REAL DATA EXAMPLE

To illustrate the modeling in a real data example, we analyze the dataset of Card (1995), which contains sample information on wages, educational attainment, and other covariates in 1976 of 3,010 men. The objective is to find the effect of education, the treatment, on wages. An answer to this question cannot in general be obtained by regressing education on wages if it is suspected that both the level of education and wages are determined by unobserved confounders (e.g., ability and motivation). In his analysis, Card argued that a dummy variable representing proximity to a four-year college was a suitable instrument on the grounds that the choice of residential location should have no direct effect on the wage rate, given covariates, but that living close to college should affect the level of educational attainment. We proceed under this assumption, but to fit the problem into our setup we convert educational attainment into a binary variable that equals one if the individual attains more than 12 years of schooling and zero otherwise. The outcome variable is the individual's hourly wage rate in 1976.

We analyze these data with our semiparametric model, letting one of the covariates, years of experience in 1976, have a nonparametric effect on both the outcome and the intake. For comparison, we also fit a parametric model in which experience is modeled by a quadratic function. Such parametric models are standard in the econometrics literature that uses instrumental variables. Specifically, in the parametric specification, the covariates in the outcome equation are a constant, experience, experience squared, and indicator variables for race, whether the individual lived in a standard metropolitan statistical area in 1976, and whether the individual lived in the South in 1976. The treatment equation contains a constant and the indicator variables for race, metropolitan area, and south, and distance from a four-year college. In the semiparametric version, we omit experience and experience squared and model the effect of experience by nonparametric functions in both the outcome and treatment equations.

Our results are based on 20,000 iterations of Algorithm 1 (p. 9), following a burn-in of a 1,000 iterations. The fitting of the quadratic model is also according to Algorithm 1 without the steps involving the nonparametric functions.

Table 3. Parametric Model: Prior and Posterior Summary

	Prior		*Posterior*			
Parameter	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*Lower 2.5%*	*Upper 97.5%*
α_1	5.000	100.000	5.7901	0.1349	5.5248	6.0417
α_2	0.000	10.000	0.0708	0.0139	0.0443	0.0982
α_3	0.000	10.000	−0.0026	0.0004	−0.0035	−0.0017
α_4	0.000	10.000	−0.2391	0.0252	−0.2878	−0.1895
α_5	0.000	10.000	0.1847	0.0192	0.1470	0.2223
α_6	0.000	10.000	−0.1516	0.0158	−0.1826	−0.1205
β	0.100	100.000	0.1547	0.1066	-0.0419	0.3628
γ_1	0.000	10.000	2.5745	0.1844	2.2143	2.9382
γ_2	0.000	10.000	−0.4258	0.0364	−0.4966	−0.3543
γ_3	0.000	10.000	0.0108	0.0018	0.0072	0.0142
γ_4	0.000	10.000	−0.5844	0.0673	−0.7182	−0.4514
γ_5	0.000	10.000	0.2822	0.0640	0.1579	0.4094
γ_6	0.000	10.000	0.0391	0.0588	−0.0750	0.1562
δ	0.000	10.000	0.1992	0.0633	0.0735	0.3229
ω_{11}	5.095	10.722	0.1548	0.0054	0.1457	0.1672
ω_{12}	0.150	1.485	0.0480	0.0634	−0.0766	0.1645

Table 4. Nonparametric Model: Prior and Posterior Summary

	Prior		*Posterior*			
Parameter	*Mean*	*Std. dev.*	*Mean*	*Std. dev.*	*Lower 2.5%*	*Upper 97.5%*
α_1	5.000	100.000	5.9138	0.1489	5.6446	6.2248
α_2	0.000	10.000	−0.2162	0.0247	−0.2654	−0.1691
α_3	0.000	10.000	0.1728	0.0184	0.1368	0.2094
α_4	0.000	10.000	−0.1507	0.0159	−0.1818	−0.1199
β	0.100	100.000	0.2911	0.1052	0.0744	0.4719
γ_1	5.000	10.000	4.3880	1.7946	1.0612	7.6988
γ_2	0.000	10.000	−0.5857	0.0722	−0.7288	−0.4445
γ_3	0.000	10.000	0.2285	0.0675	0.0979	0.3601
γ_4	0.000	10.000	0.0414	0.0619	−0.0792	0.1613
δ	0.000	10.000	0.2298	0.0659	0.1009	0.3580
ω_{11}	50.925	122.140	0.1532	0.0047	0.1447	0.1634
ω_{12}	0.150	14.788	−0.0330	0.0647	−0.1439	0.0999

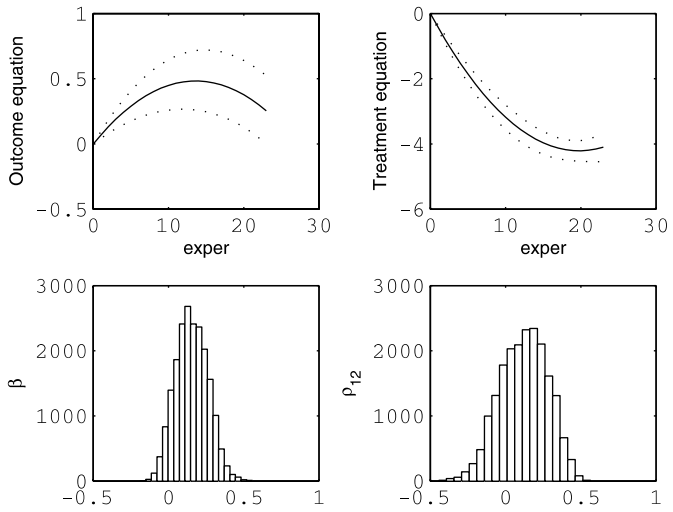

Figure 10. Parametric Model: (top panel) estimated effect of experience in the outcome and treatment models (bottom panel) marginal posterior distributions of β, and $\rho_{12} = \omega_{12}/\sqrt{\omega_{11}}$.

The marginal likelihood estimates strongly favor the semiparametric model; its base 10 logarithm is -1262.6967, compared to -1304.8112 for the parametric model. In Tables 3 and 4 we report the prior-posterior summary of the model parameters from each model. Although the results are broadly similar (e.g., the posterior credibility interval of the correlation between the errors in each case includes zero, suggesting the absence of unobserved confounders, given the covariates) there are some key differences. For one, the relationship between experience and wages is not the same between the two models. This relationship is similar across models in the outcome model, as can be seen from Figures 10 and 11, which also contain the marginal posterior distributions of β and the correlation $\rho_{12} = \omega_{12}/\sqrt{\omega_{11}}$, but this relationship is not as close in the intake equation. Another difference is in the inference about the key parameter β. The estimate of this parameter is considerably larger from the nonparametric model. The respective posterior distributions given in Figures 10 and 11 are similar in shape but the one from the nonparametric fitting is shifted to the right with a mean value of 0.2911 compared to 0.1547 from the parametric model.

7. CONCLUSIONS

We have shown how Bayesian methods may be used to analyze a semiparametric IV model. Our experiments have shown that our method is capable of recovering the shapes of highly nonlinear functions even in models with as many as three unknown functions. We have also discussed the computation of the marginal likelihood and showed that it provides a useful approach for comparing alternative specifications.

Although estimation of the nonlinear functions proved to be robust with respect to number of functions, sample size, and the value of ω_{12}, estimation of the parameters β and ω_{12} was less so when the outcome variable was binary. It appears that large sample

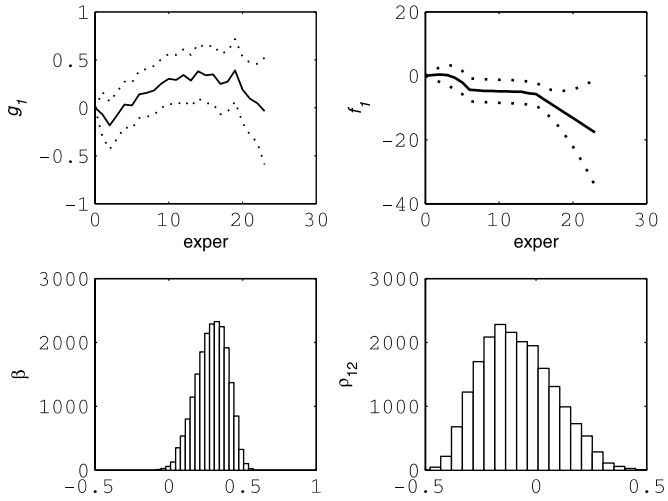

Figure 11. Nonparametric Model: (top panel) estimated effect of experience in the outcome and treatment models (bottom panel) marginal posterior distributions of β, and $\rho_{12} = \omega_{12}/\sqrt{\omega_{11}}$.

sizes are required in this case, especially when there is reason to suspect a high degree of confounding.

APPENDIX: ALGORITHM 1. SAMPLING OF THE PARAMETERS AND LATENT VARIABLES

A.1 SAMPLING $\boldsymbol{\tau}^2$ AND a

From the prior for τ_j^2 in (2.11) and the conditional prior densities of (2.8) and (2.9), it is seen that

$$\pi(\tau_j^2|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\boldsymbol{\tau}^2) = \mathcal{IG}\left(\tau_j^2|\frac{\nu_{00}+m_j-1}{2}, \frac{\delta_{00}+b_j}{2}\right), \tag{A.1}$$

where $b_j = (\boldsymbol{\Delta}_j\tilde{\mathbf{g}}_j - \mathbf{g}_{j0})'\mathbf{H}_j^{-1}(\boldsymbol{\Delta}_j\tilde{\mathbf{g}}_j - \mathbf{g}_{j0})$ for the first p_1 terms and $b_j = (\boldsymbol{\Delta}_j\tilde{\mathbf{f}}_j - \mathbf{f}_{j0})'\mathbf{H}_j^{-1}(\boldsymbol{\Delta}_j\tilde{\mathbf{f}}_j - \mathbf{f}_{j0})$ for the last q_1 terms. Each τ_j^2 is drawn independently from its posterior distribution. A similar calculation shows that

$$\pi(a_j|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\mathbf{a}) = \mathcal{IG}\left(a_j|\frac{\nu_{000}+1}{2}, \frac{\tau_j^{-2}c_j+\delta_{000}}{2}\right), \tag{A.2}$$

where $c_j = (\tilde{g}_{j2} - g_{j20})^2$ for $j = 1, \ldots, p_1$ and $(\tilde{f}_{j2} - f_{j20})^2$ for $j = p_1+1, \ldots, p_1+q_1$.

A.2 SAMPLING THE LATENT VARIABLES

From (2.2) we see that x_i^*, conditional on y_i and the remaining unknowns, has a normal distribution with mean m_i and variance v_i of Equations (2.3) and (2.4), respectively. Accordingly, after conditioning on x_i, x_i^* is sampled from a $\mathcal{N}(m_i, v_i)$ distribution truncated

to the interval $(0, \infty)$ when $x_i = 1$ and truncated to the interval $(-\infty, 0]$ when $x_i = 0$. Sampling these distributions is straightforward by the inverse cdf method.

A.3 SAMPLING $\boldsymbol{\lambda} = (\boldsymbol{\alpha}', \boldsymbol{\gamma}', \delta)'$

To derive the distribution of $\boldsymbol{\lambda}$ given the data and $\boldsymbol{\psi}\backslash\boldsymbol{\lambda}$, note that the joint conditional density of $\hat{\mathbf{y}}_i = (y_i - g(\mathbf{v}_{1i}) - x_i\beta, x_i^* - f(\mathbf{w}_{1i}))'$ is proportional to

$$\exp\left(-\frac{1}{2}(\hat{\mathbf{y}}_i - \mathbf{X}_i\boldsymbol{\lambda})'\boldsymbol{\Omega}^{-1}(\hat{\mathbf{y}}_i - \mathbf{X}_i\boldsymbol{\lambda})\right),$$

where

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{v}_{0i}' & \mathbf{0}' & \mathbf{0}' \\ \mathbf{0}' & \mathbf{w}_{0i}' & z_i \end{pmatrix}.$$

On combining this density with the prior of $\boldsymbol{\lambda}$ we have

$$\pi(\boldsymbol{\lambda}|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\boldsymbol{\lambda}) = \mathcal{N}(\boldsymbol{\lambda}|\hat{\boldsymbol{\lambda}}, \mathbf{L}), \tag{A.3}$$

where $\mathbf{L} = (\mathbf{L}_0^{-1} + \sum_i \mathbf{X}_i'\boldsymbol{\Omega}^{-1}\mathbf{X}_i)^{-1}$ and $\hat{\boldsymbol{\lambda}} = \boldsymbol{L}(\mathbf{L}_0^{-1}\boldsymbol{\lambda}_0 + \sum_i \mathbf{X}_i'\boldsymbol{\Omega}^{-1}\hat{\mathbf{y}}_i)$.

A.4 SAMPLING β, σ_{11}, AND ω_{12}:

These parameters are sampled jointly by the method of composition, first sampling σ_{11} marginalized over $\boldsymbol{\beta} = (\beta, \omega_{12})'$ and then sampling $\boldsymbol{\beta}$ given σ_{11}. For the first of these steps, rewrite the model to isolate $\boldsymbol{\beta}$. Upon defining

$$\hat{\mathbf{y}} = \mathbf{y} - \mathbf{V}_0\boldsymbol{\alpha} - \sum_k \mathbf{Q}_k\tilde{\mathbf{g}}_k$$

$$\hat{\mathbf{x}}^* = \mathbf{x}^* - \mathbf{W}_0\boldsymbol{\gamma} - \sum_k \mathbf{Q}_k\tilde{\mathbf{f}}_k - \mathbf{z}\delta,$$

we have that $\hat{\mathbf{y}}|\boldsymbol{\psi} \sim \mathcal{N}_n(\mathbf{X}^*\boldsymbol{\beta}, \sigma_{11}\mathbf{I}_n)$, where $\mathbf{X}^* = (\mathbf{x}, \hat{\mathbf{x}}^*)$, which implies that the latter distribution marginalized over the prior of $\boldsymbol{\beta}$ is normal with mean $\mathbf{X}^*\mathbf{b}_0$ and variance $\sigma_{11}(\mathbf{I}_n + \mathbf{X}^*\mathbf{B}_0\mathbf{X}^{*\prime})$. Combining this marginalized distribution with the inverse gamma prior of σ_{11} leads to the updated distribution

$$\sigma_{11}|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash(\sigma_{11}, \boldsymbol{\beta}) \sim \mathcal{IG}\left(\sigma_{11}|\frac{\nu_0 + n}{2}, \frac{\delta_0 + d}{2}\right), \tag{A.4}$$

where $d = (\hat{\mathbf{y}} - \mathbf{X}^*\mathbf{b}_0)'(\mathbf{I}_n + \mathbf{X}^*\mathbf{B}_0\mathbf{X}^{*\prime})^{-1}(\hat{\mathbf{y}} - \mathbf{X}^*\mathbf{b}_0)$. The conditional distribution of $\boldsymbol{\beta}$ follows from the distribution of $\hat{\mathbf{y}}|\boldsymbol{\psi}$ given above and the normal prior of $\boldsymbol{\beta}$; it is

$$\boldsymbol{\beta}|\mathbf{y}, \mathbf{x}, \boldsymbol{\psi}\backslash\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{b}_1, \sigma_{11}\mathbf{B}_1), \tag{A.5}$$

where $\mathbf{B}_1 = (\mathbf{B}_0^{-1} + \mathbf{X}^{*'}\mathbf{X}^*)^{-1}$ and $\mathbf{b}_1 = \mathbf{B}_1(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{X}^{*\prime}\hat{\mathbf{y}})$.

ACKNOWLEDGMENTS

Many thanks to Robert Parks for his valuable help and to Brian Gunia for his excellent research assistance. We are also grateful for the comments of two referees and an associate editor.

[Received February 2005. Revised June 2006.]

REFERENCES

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

Baladandayuthapani, V., Mallick, B. K., and Carroll, R. J. (2005), "Spatially Adaptive Bayesian Penalized Regression Splines (P-splines)," *Journal of Computational and Graphical Statistics*, 14, 378–394.

Basu, S., and Chib, S. (2003), "Marginal Likelihood and Bayes Factors for Dirichlet Process Mixture Models," *Journal of the American Statistical Association*, 98, 224–235.

Besag, J., Green, P., Higdon, D., and Mengersen, K. (1995), "Bayesian Computation and Stochastic Systems" (with discussion), *Statistical Science*, 10, 3–66.

Biller, C. (2000), "Adaptive Bayesian Regression Splines in Semiparamtric Generalized Linear Models," *Journal of Computational and Graphical Statistics*, 9, 122–140.

Card, D. (1995), "Using Geographic Variation in College Proximity to Estimate the Returns to Schooling," in *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, eds. Christophides, L. N., Grant, E. K., and Swidinsky, R., Toronto: University of Toronto Press, pp. 201–222.

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

——— (2003), "On Inferring Effects of Binary Treatments with Unobserved Confounders" (with discussion), in *Bayesian Statistics 7*, eds. J. Bernardo, M. Bayarri, J. Berger, A. Dawid, D. Heckerman, A. Smith, and M. West, London: Oxford University Press, pp. 66–84.

Chib, S., and Greenberg, E. (1995), "Understanding the Metropolis-Hastings Algorithm," *The American Statistician*, 49, 327–335.

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

Chib, S., and Jeliazkov, I. (2001), "Marginal Likelihood from the Metropolis-Hastings Output," *Journal of the American Statistical Association*, 96, 270–281.

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

Choudhuri, N., Ghosal, S., and Roy, A. (2003), "Nonparametric Binary Regression Using a Gaussian Process Prior," Working Paper, North Carolina State University Department of Statistics.

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

Efromovich, S. (1999), *Nonparametric Curve Estimation: Methods, Theory, and Applications*, Springer Series in Statistics, New York: Springer.

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.

Gersovitz, M., and MacKinnon, J. (1978), "Seasonality in Regression: An Application of Smoothness Priors," *Journal of the American Statistical Association*, 73, 264–273.

Greenland, S. (2000), "An Introduction to Instrumental Variables for Epidemiologists," *International Journal of Epidemiology*, 29, 722–729.

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

Holmes, C. C., and Mallick, B. K. (2001), "Generalized Nonlinear Modeling With Multivariate Free-Knot Regression Splines," *Journal of the Royal Statistical Society*, Ser. B, 63, 3–17.

——— (2003), "Bayesian Regression with Multivariate Linear Splines," *Journal of the American Statistical Association*, 98, 352–368.

Jeffreys, H. (1961), *Theory of Probability* (3rd ed.), Oxford: Oxford University Press.

Koop, G., and Poirier, D. J. (2004), "Bayesian Variants of Some Classical Semiparametric Regression Techniques," *Journal of Econometrics*, in press.

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

McClellan, M., McNeil, B. J., and Newhouse, J. P. (1994), "Does More Intensive Treatment of Acute Myocardial Infarction in the Elderly Reduce Mortality? Analysis Using Instrumental Variables," *Journal of the American Medical Association*, 272, 859–866.

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.

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.

Smith, M., and Kohn, R. (2000), "Nonparametric Seemingly Unrelated Regression," *Journal of Econometrics*, 98, 257–281.

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

Yatchew, A. (2003), *Semiparametric Regression for the Applied Econometrician*, Cambridge: Cambridge University Press.