Inference in Semiparametric Dynamic Models for Binary Longitudinal Data

Siddhartha CHIB and Ivan JELIAZKOV

This article deals with the analysis of a hierarchical semiparametric model for dynamic binary longitudinal responses. The main complicating components of the model are an unknown covariate function and serial correlation in the errors. Existing estimation methods for models with these features are of $\mathcal{O}(N^3)$, where N is the total number of observations in the sample. Therefore, nonparametric estimation is largely infeasible when the sample size is large, as in typical in the longitudinal setting. Here we propose a new $\mathcal{O}(N)$ Markov chain Monte Carlo based algorithm for estimation of the nonparametric function when the errors are correlated, thus contributing to the growing literature on semiparametric and nonparametric mixed-effects models for binary data. In addition, we address the problem of model choice to enable the formal comparison of our semiparametric model with competing parametric and semiparametric specifications. The performance of the methods is illustrated with detailed studies involving simulated and real data.

KEY WORDS: Average covariate effect; Bayes factor; Bayesian model comparison; Clustered data; Correlated binary data; Labor force participation; Longitudinal data; Marginal likelihood; Markov chain Monte Carlo; Markov process priors; Nonparametric estimation; Partially linear model.

1. INTRODUCTION

This article discusses techniques for analyzing semiparametric models for dynamic binary longitudinal data. A hierarchical Bayesian approach is adopted to combine the main regression components: nonparametric functional form, dynamic dependence through lags of the response variable and serial correlation in the errors, and multidimensional heterogeneity. In this context, we pursue several objectives. First, we propose new computationally efficient estimation techniques to carry out the analysis. Computational efficiency is key, because longitudinal (panel) data pose special computational challenges compared with cross-sectional or time series data, so that "brute force" estimation becomes infeasible in many settings. Second, we address the problem of model choice by computing marginal likelihoods and Bayes factors to determine the posterior probabilities of competing models. This allows for the formal comparison of semiparametric versus parametric models, and addresses the problems of variable selection and lag determination. Third, we propose a simulation-based approach to calculate the average covariate effects, which provides interpretability of the estimates despite the nonlinearity and intertemporal dependence in the model. We examine the methods in a simulation study and then apply them in the analysis of women's labor force participation.

To illustrate the model, let y_{it} be the binary response of interest, where the indices i and t ($i = 1, \ldots, n$, $t = 1, \ldots, T_i$) refer to units (e.g., individuals, firms, countries) and time. We consider a dynamic partially linear binary choice model where y_{it} depends parametrically on the covariate vectors $\tilde{\mathbf{x}}_{it}$ and $\mathbf{w}_{it}$ (containing two disjoint sets of covariates) and nonparametrically on the covariate s_{it} in the form

$$y_{it} = \mathbb{1}\big\{\tilde{\mathbf{x}}_{it}'\boldsymbol{\delta} + \mathbf{w}_{it}'\boldsymbol{\beta}_i + g(s_{it}) + \phi_1 y_{i,t-1} + \cdots + \phi_J y_{i,t-J} + \varepsilon_{it} > 0\big\}, \quad (1)$$

where $\mathbb{1}\{\cdot\}$ is an indicator function, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}_i$ are vectors of common (fixed) and unit-specific (random) effects, $\phi_1, \ldots, \phi_J$ are lag coefficients, $g(\cdot)$ is an unknown function, and ε_{it} is a serially correlated error term. We assume that $g(\cdot)$ is a smooth but otherwise unrestricted function that is estimated nonparametrically. For this reason, $\tilde{\mathbf{x}}_{it}$ does not contain an intercept or the covariate s_{it}, although those may be included in $\mathbf{w}_{it}$; identification issues are discussed in Section 2.2. The model in (1) aims to exploit the panel structure to distinguish among three important sources of intertemporal dependence in the observations $(y_{i1}, \ldots, y_{iT_i})$. One source is due to the lags $y_{i,t-1}, \ldots, y_{i,t-J}$, which capture the notion of "state dependence," where the probability of response may depend on past occurrences because of altered preferences, trade-offs, or constraints. A second source of dependence is the presence of serial correlation in the errors $(\varepsilon_{i1}, \ldots, \varepsilon_{iT_i})$. Finally, the observations $(y_{i1}, \ldots, y_{iT_i})$ can also be correlated because of heterogeneity; these differences are captured through the individual effects $\boldsymbol{\beta}_i$. Addressing the differences among units is also essential in guarding against the emergence of "spurious state dependence," because temporal pseudodependence could occur due to the fact that history may simply serve as a proxy for these unobserved differences (Heckman 1981).

The presence of the nonparametric function in the binary response model in (1) raises a number of challenges for estimation, because of the intractability of the likelihood function. Many of these problems have been largely overcome in the Bayesian context by Wood and Kohn (1998) and Shively, Kohn, and Wood (1999), based on the framework of Albert and Chib (1993). One open problem, however, is the analysis of semiparametric models with serially correlated errors. This issue has been studied by Diggle and Hutchinson (1989), Altman (1990), Smith, Wong, and Kohn (1998), Wang (1998), and Opsomer, Wang, and Yang (2001). These studies conclude that serial correlation, if ignored, poses fundamental problems that can have substantial adverse consequences for estimation of the nonparametric function. The ability to estimate models with serial correlation is limited in practice, however, because of the computational intensity of existing algorithms. In these cases the estimation algorithms are $\mathcal{O}(N^3)$, where $N = \sum_{i=1}^{n} T_i$

Siddhartha Chib is the Harry C. Hartkopf Professor of Econometrics and Statistics, John M. Olin School of Business, Washington University, St. Louis, MO 63130 (E-mail: *chib@wustl.edu*). Ivan Jeliazkov is Assistant Professor, Department of Economics, University of California, Irvine, CA 92697 (E-mail: *ivan@uci.edu*). The authors thank the editor, associate editor, and two referees, along with Edward Greenberg, Dale Poirier, and Justin Tobias, for helpful comments.

Journal of the American Statistical Association
June 2006, Vol. 101, No. 474, Theory and Methods
DOI 10.1198/016214505000000871

is the total number of observations in the sample. This feature renders nonparametric estimation infeasible in panel data settings where sample sizes are generally large. Here we propose a new $\mathcal{O}(N)$ algorithm for estimating the nonparametric function when the errors are correlated, thus contributing to the growing literature on static semiparametric and nonparametric mixed-effects models for binary data (see, e.g., Lin and Zhang 1999; Lin and Carroll 2001; Karcher and Wang 2001). As far as we know, in the literature there are no $\mathcal{O}(N)$ estimators for (binary or continuous) panel data with correlated errors.

A second open problem is the question of model comparison in the semiparametric setting. Because the marginal likelihood, which is used to compare two models on the basis of their Bayes factor, is a large-dimensional integral, previous work (e.g., DiMatteo, Genovese, and Kass 2001; Wood, Kohn, Shively, and Jiang 2002; Hansen and Kooperberg 2002) has relied on measures such as the Akaike information criterion (AIC) and Bayes information criterion (BIC). However, computation of these criteria is infeasible in our context. We extend that literature by describing an approach, based on Chib (1995), for calculating marginal likelihoods and Bayes factors. We use this approach in Section 7 to compare several competing parametric and semiparametric models.

The article is organized as follows. Section 2 completes the statistical model, and Section 3 presents the Markov chain Monte Carlo (MCMC) fitting method. Section 4 shows how the average effects of the covariates on the probability of response are calculated, and Section 5 is concerned with model comparison. Section 6 presents a detailed simulation study of the performance of the estimation method. Section 7 studies the intertemporal labor force participation of a panel of married women. Finally, Section 8 presents concluding remarks.

2. HIERARCHICAL MODELING AND PRIORS

This section presents the hierarchical structure used in modeling the regression components in (1). For simplicity, and to keep the discussion focused on our main topics—semiparametric estimation with correlated errors and Bayesian model comparison—we present the model in detail only under the assumptions of this section. A number of modifications, extensions, and generalizations are possible, and we mention some of them as we proceed.

2.1 The Smoothness Prior on $g(\cdot)$

Suppose that the N observations in the covariate vector $\mathbf{s}$, whose effect is modeled nonparametrically, determine the $m \times 1$ *design point vector* $\mathbf{v}$, $m \leq N$, with entries equal to the *unique ordered* values of $\mathbf{s}$ with $v_1 < \cdots < v_m$ and with $\mathbf{g} = (g(v_1), \ldots, g(v_m))' = (g_1, \ldots, g_m)'$ as the corresponding function evaluations. The idea is to model the function evaluations as a stochastic process that controls the degree of local variation between neighboring states in $\mathbf{g}$ to strike a balance between a good fit and a smooth regression function (Whittaker 1923). Following Fahrmeir and Lang (2001), we model the function evaluations as resulting from the realization of a second-order Markov process, with the specification aimed at penalizing rough functions $g(\cdot)$. A range of similar smoothness priors, with some discussion and comparisons, can be found in the literature on nonparametric modeling (for some specific examples, see Wahba 1978; Shiller 1984; Silverman 1985; Besag, Green, Higdon, and Mengersen 1995; Koop and Poirier 2004).

Defining $h_t = v_t - v_{t-1}$, the second-order random-walk specification is given by

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

where τ^2 is a smoothness parameter. Small values of τ^2 produce smoother functions, whereas larger values allow the function to be more flexible and interpolate the data more closely. The weight h_t adjusts the variance to account for possibly irregular spacing between consecutive points in the design point vector. Other possibilities are conceivable for the weights (see, e.g., Shiller 1984; Besag et al. 1995; Fahrmeir and Lang 2001); the one given here implies that the variance grows linearly with the distance h_t, a property satisfied by random walks. This linearity is appealing because it implies that conditional on g_{t-1} and g_{t-2}, the variance of g_{t+k}, $k \geq 0$, will depend only on the distance $v_{t+k} - v_{t-1}$, but not on the number of points k that lie in between.

We now deviate from the prior of Fahrmeir and Lang (2001), and, in fact, from much of the literature on nonparametric functional modeling, by working with a version of the prior in (2) that is proper. The prior in (2) is improper because it incorporates information only about deviations from linearity, but says nothing about the linearity itself. To rectify this problem, we complete the specification of the smoothness prior by providing a distribution for the initial states of the random-walk process,

$$\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} \Big| \tau^2 \sim \mathcal{N}\left(\begin{pmatrix} g_{10} \\ g_{20} \end{pmatrix}, \tau^2 \mathbf{G}_0 \right), \tag{3}$$

where $\mathbf{G}_0$ is a 2×2 symmetric positive definite matrix. The prior on the initial conditions (3) induces a prior on linear functions of $\mathbf{v}$ that is equivalent to the usual priors placed on the intercept and slope parameters in univariate linear regression. This can be seen more precisely by iterating (2) in expectation (to eliminate u_t which is the source of the nonlinearity), starting with initial states as specified in (3). Thus, conditional on g_1 and g_2, the mean of the Markov process in (2) is a straight line that goes through g_1 and g_2. As a consequence, the intercept and slope of that line will have a distribution directly related to the distribution in (3) in a one-to-one mapping. This is useful in setting the prior parameters g_{10}, g_{20}, and $\mathbf{G}_0$ based on the same information that would be used in a corresponding linear model.

The directed Markovian structure in (2) and (3) implies a joint density for the elements of $\mathbf{g}$, which can be obtained by rewriting the Markov process in a random field form. After defining

$$\mathbf{H} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ \frac{h_3}{h_2} & -(1 + \frac{h_3}{h_2}) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_m}{h_{m-1}} & -(1 + \frac{h_m}{h_{m-1}}) & 1 \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} \mathbf{G}_0 & & & \\ & h_3 & & \\ & & \ddots & \\ & & & h_m \end{pmatrix},$$

where the off-diagonal 0's in $\mathbf{H}$ and $\boldsymbol{\Sigma}$ have been suppressed in the notation, the prior on $\mathbf{g}$, which is equivalent to the second-order Markov process prior in (2) and (3), becomes

$$\mathbf{g}|\tau^2 \sim \mathcal{N}(\mathbf{g}_0, \tau^2\mathbf{K}^{-1}), \qquad (4)$$

where $\mathbf{g}_0 = \mathbf{H}^{-1}\tilde{\mathbf{g}}$, with $\tilde{\mathbf{g}} = (g_{10}, g_{20}, 0, \dots, 0)'$, and the *penalty matrix* $\mathbf{K}$ is given by $\mathbf{K} = \mathbf{H}'\boldsymbol{\Sigma}^{-1}\mathbf{H}$. Equivalently, $\mathbf{g}_0$ can be derived by taking recursive expectations of (2) starting with the mean in (3), and as argued earlier, the points in $\mathbf{g}_0$ will form a straight line.

Several points deserve emphasis. First, a key feature of the prior in (4), due to the information in (3), is that it is proper. In contrast, priors in the literature are generally specified with a reduced-rank penalty matrix $\mathbf{K}$ and thus are improper. Because improper priors preclude the possibility of formal finite-sample model comparison using marginal likelihoods and Bayes factors (O'Hagan 1994, chap. 7; Kass and Raftery 1995), our prior removes an important impediment to Bayesian model selection. Second, it is important to note that the $m \times m$ penalty matrix $\mathbf{K}$ is banded, which has considerable practical value because manipulations involving banded matrices take $\mathcal{O}(m)$ operations, rather than $\mathcal{O}(m^3)$ for inversions or $\mathcal{O}(m^2)$ for multiplication by a vector (and m may be as large as the total number of data points N). Third, Markov process priors are conceptually simple and adaptable to different orders, so as to meet problem-specific tasks (Besag et al. 1995; Fahrmeir and Lang 2001). For example, a first-order prior $g_t = g_{t-1} + u_t$ penalizes abrupt jumps between successive states of the Markov process, whereas higher-order priors embody more subtle notions of "smoothness" related to the rates of change in the function.

Because the prior on $\mathbf{g}$ is defined conditional on the hyperparameter τ^2, in the next level of the hierarchy we specify the prior distribution $\tau^2 \sim \mathcal{IG}(\nu_0/2, \delta_0/2)$. In setting the parameters ν_0 and δ_0, it is helpful to use the well-known mapping between the mean and variance of the inverse gamma distribution and the parameters ν_0 and δ_0 (e.g., Gelman, Carlin, Stern, and Rubin 1995, app. A). The choice of these parameters will affect the estimated $\mathbf{g}$ depending on the other sources of variance in the model. Some intuition can be gained by considering the sampling algorithm that we present in the next section, where in the sampling of $\mathbf{g}$, the inverse of τ^2 weighs the components of the smoothness prior $\mathbf{K}$ and $\mathbf{g}_0$ and competes with the inverse of the error variance, which weighs the function $\mathbf{g}$ that maximizes the fit in the likelihood.

2.2 Priors on the Linear Effects and Dynamic Parameters

Turning attention to the parametric facets of the model, and in anticipation of the subsequent estimation of the model by the approach of Albert and Chib (1993), we begin by rewriting the model in (1) in terms of the latent variables $\{z_{it}\}$ as

$$z_{it} = \tilde{\mathbf{x}}_{it}'\boldsymbol{\delta} + \mathbf{w}_{it}'\boldsymbol{\beta}_i + g(s_{it}) + \phi_1\mathbb{1}\{z_{i,t-1} > 0\} + \dots + \phi_J\mathbb{1}\{z_{i,t-J} > 0\} + \varepsilon_{it},$$

where $y_{it} = \mathbb{1}\{z_{it} > 0\}$. [In the presample $(t = -J+1, \dots, 0)$, the latent data $\{z_{it}\}$ are not modeled and for our purposes can simply be taken to equal the presample $\{y_{it}\}$.] Suppose that ε_{it} is a serially correlated error term that follows a mean-0 stationary pth-order autoregressive [AR(p)] process parameterized in terms of $\boldsymbol{\rho} = (\rho_1, \dots, \rho_p)'$, that is,

$$\varepsilon_{it} = \rho_1\varepsilon_{i,t-1} + \dots + \rho_p\varepsilon_{i,t-p} + \nu_{it}, \qquad (5)$$

where ν_{it} are independent $\mathcal{N}(0, 1)$. We can express the process in (5) in terms of a polynomial in the backshift operator L as $\rho(L)\varepsilon_{it} = \nu_{it}$, where $\rho(L) = 1 - \rho_1 L - \dots - \rho_p L^p$ and stationarity is maintained by requiring that all roots of $\rho(L)$ lie outside the unit circle. Estimation with this and several other correlation structures is discussed in Section 3.

Stacking the data for each cluster, let $\mathbf{y}_i \equiv (y_{i1}, \dots, y_{iT_i})'$ denote the T_i observations in the ith cluster, and similarly define the lag vectors

$$\mathbf{y}_{i,-j} \equiv (y_{i,1-j}, \dots, y_{i,T_i-j})' = (\mathbb{1}\{z_{i,1-j} > 0\}, \dots, \mathbb{1}\{z_{i,T_i-j} > 0\})', \qquad j = 1, \dots, J.$$

Then, for the observations in the ith cluster, we have that

$$\mathbf{z}_i = \tilde{\mathbf{X}}_i\boldsymbol{\delta} + \mathbf{W}_i\boldsymbol{\beta}_i + \mathbf{g}_i + \mathbf{L}_i\boldsymbol{\phi} + \boldsymbol{\varepsilon}_i, \qquad (6)$$

where $\mathbf{z}_i = (z_{i1}, \dots, z_{iT_i})'$, $\tilde{\mathbf{X}}_i = (\tilde{\mathbf{x}}_{i1}, \dots, \tilde{\mathbf{x}}_{iT_i})'$, $\mathbf{W}_i = (\mathbf{w}_{i1}, \dots, \mathbf{w}_{iT_i})'$, $\mathbf{g}_i = (g(s_{i1}), \dots, g(s_{iT_i}))'$, $\mathbf{s}_i = (s_{i1}, \dots, s_{iT_i})'$, $\mathbf{L}_i = (\mathbf{y}_{i,-1}, \dots, \mathbf{y}_{i,-J})$, $\boldsymbol{\phi} = (\phi_1, \dots, \phi_J)'$, and the errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT_i})'$ follow the distribution $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_i)$. The covariance matrix $\boldsymbol{\Omega}_i$ is the $T_i \times T_i$ Toeplitz matrix implied by the AR process, the construction of which we will discuss later in this article. But first consider the modeling of the unobserved effects.

In the spirit of Mundlak (1978), we assume that the distribution of the q-vector $\boldsymbol{\beta}_i$ is Gaussian with mean value depending on the initial observations $\mathbf{y}_{i0} \equiv (y_{i,-J+1}, \dots, y_{i0})'$ and the covariates for subject i. In particular, we let

$$\boldsymbol{\beta}_i|\mathbf{y}_{i0}, \tilde{\mathbf{X}}_i, \mathbf{W}_i, \mathbf{s}_i, \boldsymbol{\gamma}, \mathbf{D} \sim \mathcal{N}(\mathbf{A}_i\boldsymbol{\gamma}, \mathbf{D}), \qquad i = 1, \dots, n, \qquad (7)$$

or equivalently,

$$\boldsymbol{\beta}_i = \mathbf{A}_i\boldsymbol{\gamma} + \mathbf{b}_i, \qquad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \qquad i = 1, \dots, n, \qquad (8)$$

where the matrix $\mathbf{A}_i$ can be defined quite flexibly, given the specifics of the problem at hand.

In the simplest case in which $\mathbf{W}_i$ does not include an intercept and $\boldsymbol{\beta}_i$ is independent of the covariates, a parsimonious way to model the dependence of $\boldsymbol{\beta}_i$ on $\mathbf{y}_{i0}$ is to let $\mathbf{A}_i$ be a $q \times 2q$ matrix given by $\mathbf{A}_i = \mathbf{I} \otimes (1, \bar{y}_{i0})$, where $\bar{y}_{i0}$ is the mean of the entries in $\mathbf{y}_{i0}$. [More generally, but less parsimoniously, $\mathbf{A}_i$ can also be given by $\mathbf{I} \otimes (1, \mathbf{y}_{i0}')$.] Moreover, $\mathbf{A}_i$ also may contain within-cluster means (or entire covariate sequences) of a subset of covariates—those suspected of being correlated with the random effects for each cluster. If $\bar{\mathbf{r}}_{ij}$ $(j = 1, \dots, q)$ denotes the vector of such covariate means (or entire sequence of covariates), then $\mathbf{A}_i$ may be written as

$$\mathbf{A}_i = \begin{pmatrix} 1 \quad \bar{y}_{i0} \quad \bar{\mathbf{r}}_{i1}' & & \\ & \ddots & \\ & & 1 \quad \bar{y}_{i0} \quad \bar{\mathbf{r}}_{iq}' \end{pmatrix}.$$

We later adjust this specification for the case when a random intercept is present. An example of the matrix $\mathbf{A}_i$ is illustrated in the application of Section 7.

Using (8), equation (6) can be equivalently expressed as

$$\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{g}_i + \mathbf{W}_i\mathbf{b}_i + \boldsymbol{\varepsilon}_i, \tag{9}$$

with

$$\mathbf{X}_i = (\tilde{\mathbf{X}}_i \quad \mathbf{W}_i\mathbf{A}_i \quad \mathbf{L}_i) \qquad \text{and} \qquad \boldsymbol{\beta} = (\boldsymbol{\delta}' \quad \boldsymbol{\gamma}' \quad \boldsymbol{\phi}')'.$$

The presence of the individual effects, $\mathbf{b}_i$, induces correlation among the observations in a cluster, but the clusters are modeled as independent. Therefore, even though the columns of $\mathbf{W}_i$ can form a subset of the columns of $\mathbf{X}_i$ (as is usual in parametric longitudinal models) and $\mathbf{W}_i$ can contain the covariate s whose effect is modeled through $g(s)$, the parameters of the model are likelihood-identified through the presence of intercluster and intracluster sources of variation. However, there is an important caveat for the semiparametric model considered here. We note that neither $\mathbf{s}_i$ nor an intercept should be present in the matrix $\mathbf{X}_i$, even though they can be in order to resolve the identification problem that would otherwise emerge under a general, unrestricted $g(s)$. Hence, in more general specifications containing a random intercept in the model, one must adjust $\mathbf{A}_i$ for identification purposes. If the random intercept is the ith column of $\mathbf{W}_i$, then the column of $\mathbf{A}_i$ that is an ith unit vector should be dropped, so that $\mathbf{W}_i\mathbf{A}_i$ does not include an intercept. Similarly, if $\mathbf{s}_i$ is the jth column of $\mathbf{W}_i$, then the column of $\mathbf{A}_i$ that is a jth unit vector should be dropped, so that the product $\mathbf{W}_i\mathbf{A}_i$ does not contain $\mathbf{s}_i$. It should also be noted that the presence of an unrestricted $g(\cdot)$ does not prevent the inclusion of temporally invariant covariates (e.g., gender, race, various dummies) in either $\tilde{\mathbf{X}}_i$ or $\mathbf{W}_i$, as long as these vary among clusters. One should be aware, however, that their simultaneous inclusion into $\mathbf{A}_i$ to model correlation with a random intercept leaves the likelihood unidentified (because $\mathbf{W}_i\mathbf{A}_i$ will cause $\mathbf{X}_i$ to contain two or more identical columns across all i).

The hierarchical structure of the model is completed by the introduction of (semiconjugate) prior densities for the model parameters $\boldsymbol{\beta}$, $\boldsymbol{\rho}$, and $\mathbf{D}$. Gaussian priors are used to summarize the prior information about the k-vector $\boldsymbol{\beta}$, whereas a Wishart prior is used for the $q \times q$ matrix $\mathbf{D}^{-1}$, namely $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\beta}_0, \mathbf{B}_0)$ and $\mathbf{D}^{-1} \sim \mathcal{W}(r_0, \mathbf{R}_0)$. The prior on $\boldsymbol{\rho}$ is specified as $\boldsymbol{\rho} \sim \mathcal{N}(\boldsymbol{\rho}_0, \mathbf{P}_0)I_{S_\rho}$, where I_{S_ρ} is an indicator of the set S_ρ containing the $\boldsymbol{\rho}$ that satisfy stationarity. We clarify that, in contrast with the serially correlated errors, stationarity is not an issue for the state-dependence coefficients $\boldsymbol{\phi}$ (which are part of $\boldsymbol{\beta}$), because these multiply the binary lags and thus serve simply as intercept shifts. The foregoing semiconjugate priors are useful for computational reasons, but the analysis of models with other general priors can be conducted by weighted resampling of the MCMC draws obtained from a model using the priors mentioned here.

3. ESTIMATION

This section presents a new estimation method for longitudinal models with unobserved heterogeneity and serial correlation in the errors. As far as we know, other $\mathcal{O}(N)$ methods for estimating this model are not yet available. A main difficulty is that even a single evaluation of the likelihood function requires evaluation of a multiple integral, and hence penalized likelihood estimation is impractical. Moreover, because m, the number of unique values of $\mathbf{s}$, can be as large as the sample size N, any maximization over the values $\mathbf{g}$ can be extremely difficult to apply. Of course, any such maximization needs to be conditioned on a smoothness parameter, but determining that parameter is problematic by current methods such as cross-validation, which are designed for continuous data, but not for the case of longitudinal binary data.

Our estimation algorithm takes advantage of the approach of Albert and Chib (1993) to simplify simulation of the posterior distribution by MCMC methods. In the longitudinal data setup, the latent data representation of the model with AR(p) serial correlation was given in (9), where the errors $\boldsymbol{\varepsilon}_i = (\varepsilon_{i1}, \dots, \varepsilon_{iT_i})'$ follow the distribution $\boldsymbol{\varepsilon}_i \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}_i)$ and $\boldsymbol{\Omega}_i$ is the $T_i \times T_i$ Toeplitz matrix implied by the AR process. For the general AR(p) case, the matrix $\boldsymbol{\Omega}_i$ can be determined as follows. Let $\varphi_j = E(\varepsilon_{it}\varepsilon_{i,t-j})$ be the jth autocovariance (satisfying $\varphi_j = \varphi_{-j}$). It can be shown (cf. Hamilton 1994, sec. 3.4) that the autocovariances follow the same pth-order difference equation as the process itself, that is, $\varphi_j = \rho_1\varphi_{j-1} + \cdots + \rho_p\varphi_{j-p}$. The first p values $(\varphi_0, \varphi_1, \dots, \varphi_{p-1})$ are given by the first p elements of the first column of the $p^2 \times p^2$ matrix $[\mathbf{I} - \mathbf{F} \otimes \mathbf{F}]^{-1}$, where $\otimes$ denotes the Kronecker product and $\mathbf{F}$ is the $p \times p$ matrix

$$\mathbf{F} \equiv \begin{bmatrix} \boldsymbol{\rho}' \\ \begin{matrix} \mathbf{I}_{p-1} & \mathbf{0}_{(p-1)\times 1} \end{matrix} \end{bmatrix}.$$

Using the sequence of autocovariances φ_j obtained in this way, the matrix $\boldsymbol{\Omega}_i$ can be constructed using $\boldsymbol{\Omega}_i[j,k] = \varphi_{j-k}$. For example, in the AR(1) case, we have $\boldsymbol{\Omega}_i[j,k] = \rho^{|j-k|}/(1-\rho^2)$.

After marginalizing $\mathbf{b}_i$ using the distribution of the random effects, the latent $\mathbf{z}_i$ can be expressed as

$$\mathbf{z}_i = \mathbf{X}_i\boldsymbol{\beta} + \mathbf{g}_i + \mathbf{u}_i, \tag{10}$$

where the error vector is normal with variance matrix $\mathbf{V}_i = \boldsymbol{\Omega}_i + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$. This implies that the contribution of the ith cluster to the likelihood function (conditioned on $\mathbf{g}_i$),

$$\Pr(\mathbf{y}_i|\boldsymbol{\beta}, \mathbf{g}_i, \mathbf{D}, \boldsymbol{\rho}) = \int_{B_{iT_i}} \cdots \int_{B_{i1}} \mathcal{N}(\mathbf{z}_i|\mathbf{X}_i\boldsymbol{\beta} + \mathbf{g}_i, \mathbf{V}_i)\, d\mathbf{z}_i, \tag{11}$$

where B_{it} is the interval $(0, \infty)$ if $y_{it} = 1$ or the interval $(-\infty, 0]$ if $y_{it} = 0$ is, in general, difficult to calculate. This is true even in the case of uncorrelated errors when $\mathbf{V}_i = \mathbf{I} + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$, making it impractical to obtain smoothness parameter estimates by cross-validation and to subsequently use penalized likelihood estimation to obtain estimates of $\mathbf{g}$ in generalized mixed-effects models.

Previous studies by Diggle and Hutchinson (1989), Altman (1990), and Smith et al. (1998) have drawn attention to the fact that when the errors are treated as independent when they are not, the correlation in the errors can adversely affect the nonparametric estimate of the regression function. For example, when the covariate $\mathbf{s}$ is in temporal order, the unknown function $g(s)$ can be confounded with the autocorrelated error process, because both are stochastic processes in time. If the serial correlation in the errors is ignored, then the estimate of $g(s)$ can become too rough as it attempts to mimic the errors. The undersmoothing can be visible even for mild serial correlation.

Smith et al. (1998) pointed out that even if the independent variable is not time, modeling the autocorrelation in the errors gives more efficient nonparametric estimates, because it reduces the effective error variance similarly to the case of parametric regression.

To describe the approach, we stack the observations in (9) for all n clusters and write

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{g} + \mathbf{W}\mathbf{b} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Omega}), \tag{12}$$

after defining the vectors $\mathbf{z} = (\mathbf{z}_1', \ldots, \mathbf{z}_n')'$ and $\mathbf{b} = (\mathbf{b}_1', \ldots, \mathbf{b}_n')'$ and the matrices

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix}, \qquad \mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & \\ & \ddots & \\ & & \mathbf{W}_n \end{bmatrix}, \qquad \text{and}$$

$$\boldsymbol{\Omega} = \begin{bmatrix} \boldsymbol{\Omega}_1 & & \\ & \ddots & \\ & & \boldsymbol{\Omega}_n \end{bmatrix}.$$

In the foregoing, $\mathbf{Q}$ is an *incidence matrix* of dimension $N \times m$, with entries $\mathbf{Q}_{ij} = 1$ if $s_i = v_j$ and 0 otherwise. In other words, the ith row of $\mathbf{Q}$ contains a 1 in the position where the observation on $\mathbf{s}$ for that row matches the design point from the vector $\mathbf{v}$, with all remaining elements 0's, so that $\mathbf{s} = \mathbf{Q}\mathbf{v}$. The fact that the errors are not orthogonal (and $\boldsymbol{\Omega}$ is not diagonal) requires only minor adjustments to the sampling of $\boldsymbol{\beta}$, $\mathbf{z}$, $\mathbf{D}$, τ^2, and $\mathbf{b}$; standard Bayes updates for models with serial correlation can be applied to obtain and simulate the posterior densities (given in Algorithm 1 later in this section). But the sampling of $\mathbf{g}$ is problematic.

To understand the difficulty, note that with uncorrelated errors (i.e., when $\boldsymbol{\Omega} = \mathbf{I}$), we have the following full-conditional distribution: $\mathbf{g}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \tau^2, \mathbf{z} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$. In this case, $\mathbf{G} = (\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{Q})^{-1}$ and $\hat{\mathbf{g}} = \mathbf{G}(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{b}))$. Remark 1 presents a computationally efficient method for sampling this density.

Remark 1. In sampling $\mathbf{g}$, one should note that $\mathbf{Q}'\mathbf{Q}$ is a diagonal matrix with a jth diagonal entry equal to the number of values in $\mathbf{s}$ corresponding to the design point v_j. Because $\mathbf{K}$ and $\mathbf{Q}'\mathbf{Q}$ are banded, $\mathbf{G}^{-1}$ is banded as well. Thus sampling of $\mathbf{g}$ need not include an inversion to obtain $\mathbf{G}$ and $\hat{\mathbf{g}}$. The mean $\hat{\mathbf{g}}$ can be found instead by solving $\mathbf{G}^{-1}\hat{\mathbf{g}} = (\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{b}))$, which is done in $\mathcal{O}(m)$ operations by back substitution. Also, let $\mathbf{P}'\mathbf{P} = \mathbf{G}^{-1}$, where $\mathbf{P}$ is the Cholesky decomposition of $\mathbf{G}^{-1}$ and is also banded. To efficiently obtain a random draw from $\mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$, sample $\mathbf{u} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and solve $\mathbf{P}\mathbf{x} = \mathbf{u}$ for $\mathbf{x}$ by back substitution. It follows that $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \mathbf{G})$. Adding the mean $\hat{\mathbf{g}}$ to $\mathbf{x}$, yields a draw $\mathbf{g} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$.

Unfortunately, after accounting for the autocorrelated errors, the full-conditional distribution $\mathbf{g}|\mathbf{y}, \boldsymbol{\beta}, \mathbf{b}, \tau^2, \mathbf{z}, \boldsymbol{\Omega} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$ will involve the matrix $\mathbf{G}^{-1} = (\mathbf{K}/\tau^2 + \mathbf{Q}'\boldsymbol{\Omega}^{-1}\mathbf{Q})$, which is no longer banded (even though $\boldsymbol{\Omega}^{-1}$ is banded). Hence the computational shortcuts discussed in Remark 1 are inapplicable. Intuitively, bandedness fails because serial correlation introduces dependence between observations that are neighbors on the basis of the ordering of the covariate $\mathbf{s}$, whereas the function evaluations $\mathbf{g}$ depend on neighbors determined according to the ordering in $\mathbf{v}$, the vector of unique and ordered values of $\mathbf{s}$ (with $\mathbf{s} = \mathbf{Q}\mathbf{v}$).

Diggle and Hutchinson (1989) and Altman (1990) considered a special case that can still result in $\mathcal{O}(N)$ estimation. In this case, attention was restricted to univariate models for nonclustered data where the independent variable is time. Then, because the elements in $\mathbf{s}$ are already unique and ordered, we have $\mathbf{v} = \mathbf{s}$ or, in other words, $\mathbf{Q} = \mathbf{I}$. This implies that $\mathbf{G}^{-1} = (\mathbf{K}/\tau^2 + \mathbf{Q}'\boldsymbol{\Omega}^{-1}\mathbf{Q}) = (\mathbf{K}/\tau^2 + \boldsymbol{\Omega}^{-1})$ is banded, and estimation can be done in $\mathcal{O}(N)$ operations as outlined in Remark 1. Unfortunately, in panel data settings $\mathbf{Q}$ is unlikely to be an identity matrix even when $\mathbf{s}$ is time, because repeating values in $\mathbf{s}$ will tend to emerge across clusters. Even when all values in $\mathbf{s}$ are unique, $\mathbf{Q}$ will be some permutation (not necessarily an identity) matrix, because the ordering of $\mathbf{s}$ need not be consistent with the cluster structure. The general case where $\mathbf{s}$ is allowed to be any covariate (not necessarily time) was considered by Smith et al. (1998), but their algorithm was $\mathcal{O}(N^3)$, which works with the nonbanded matrix $\mathbf{G}^{-1}$. Thus the applicability of that method is limited to only small datasets and is infeasible in panel data settings, where the sample size N can run into the thousands. Finally, we note that the method of orthogonalizing the errors by working with the transformed data $\rho(L)z_{it}$, $\rho(L)\mathbf{x}_{it}$, $\rho(L)g(s_{it})$, and $\rho(L)\mathbf{w}_{it}$ (cf. Harvey 1981, chap. 6; Chib 1993) works well in parametric models but is not a solution here, because it is equivalent to premultiplying the matrices $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Q}$ by the Cholesky decomposition of $\boldsymbol{\Omega}^{-1}$, which still leaves $\mathbf{G}^{-1}$ nonbanded.

Here we propose a different approach to orthogonalizing the errors that exploits the longitudinal nature of the data. In particular, the idea is to decompose the errors into a correlated and an orthogonal parts, and to deal with the correlated part in much the same way in which we deal with the random effects. Once the correlated part is given, the nonparametric estimation of $\mathbf{g}$ can proceed as efficiently as before. To illustrate, decompose the matrix $\boldsymbol{\Omega}_i = \mathbf{R}_i + \kappa\mathbf{I}$, where $\mathbf{R}_i$ is a symmetric positive definite matrix and $\kappa\mathbf{I}$ is a diagonal matrix with $\kappa > 0$. Furthermore, let $\mathbf{C}_i$ be the Cholesky decomposition of $\mathbf{R}_i$ such that $\mathbf{C}_i'\mathbf{C}_i = \mathbf{R}_i$. Then $\boldsymbol{\Omega}_i = \mathbf{C}_i'\mathbf{C}_i + \kappa\mathbf{I}$, and the model can be rewritten as

$$\begin{aligned} \mathbf{z}_i &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \mathbf{g}_i + \boldsymbol{\varepsilon}_i \\ &= \mathbf{X}_i\boldsymbol{\beta} + \mathbf{W}_i\mathbf{b}_i + \mathbf{g}_i + \mathbf{C}_i'\mathbf{u}_i + \boldsymbol{\xi}_i, \end{aligned} \tag{13}$$

where $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and $\xi_i \sim \mathcal{N}(\mathbf{0}, \kappa\mathbf{I})$ are mutually independent. Stacking the observations in (13) for all n clusters, in analogy with (12), we have

$$\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Q}\mathbf{g} + \mathbf{W}\mathbf{b} + \mathbf{C}'\mathbf{u} + \boldsymbol{\xi}, \qquad \boldsymbol{\xi} \sim \mathcal{N}(\mathbf{0}, \kappa\mathbf{I}), \tag{14}$$

where $\mathbf{u} = (\mathbf{u}_1', \ldots, \mathbf{u}_n')'$ and $\mathbf{C}$ is given by

$$\mathbf{C} = \begin{bmatrix} \mathbf{C}_1 & & \\ & \ddots & \\ & & \mathbf{C}_n \end{bmatrix}.$$

Because the covariance matrix of $\boldsymbol{\xi}$ is diagonal, conditional on $\mathbf{C}'\mathbf{u}$, we have obtained an orthogonalization of the serially correlated errors that can be used to sample $\mathbf{g}$ efficiently. It now remains to prove that a decomposition of $\boldsymbol{\Omega}_i$ into the sum of a symmetric positive definite matrix $\mathbf{R}_i$ and a (positive definite) diagonal matrix $\kappa\mathbf{I}$ exists, and to show how it can be found. The details are formalized here and rely on results given by Ortega (1987, pp. 31–32).

Theorem 1. Let the matrix $\mathbf{A}$ have eigenvalues $\{\lambda_i\}$, and let $p(\mathbf{Z}) \equiv \kappa_0\mathbf{I} + \kappa_1\mathbf{Z} + \kappa_2\mathbf{Z}^2 + \cdots + \kappa_m\mathbf{Z}^m$ be some (scalar- or matrix-valued) polynomial of degree m. Then $\{p(\lambda_i)\}$ are eigenvalues of the matrix $p(\mathbf{A})$, and $\mathbf{A}$ and $p(\mathbf{A})$ have the same eigenvectors.

The proof follows from first principles by writing the matrices in the polynomial in terms of their spectral decomposition and collecting terms. The result that we originally sought to show can now be derived using a special case of the polynomial in the theorem, that is, $p(\mathbf{Z}) = \mathbf{Z} - \kappa\mathbf{I}$.

Lemma 1. A $T_i \times T_i$ symmetric positive definite matrix $\boldsymbol{\Omega}_i$ can always be written as the sum of a symmetric positive definite matrix $\mathbf{R}_i$ and a positive definite diagonal matrix $\kappa\mathbf{I}$.

Proof. Let $p(\boldsymbol{\Omega}_i) = \boldsymbol{\Omega}_i - \kappa\mathbf{I}$, which is the definition of $\mathbf{R}_i$. Then, if the eigenvalues of $\boldsymbol{\Omega}_i$ are $\{\lambda_{ij}\}_{j=1}^{T_i}$, which are strictly positive and real because $\boldsymbol{\Omega}_i$ is symmetric and positive definite, it follows from Theorem 1 that the eigenvalues of $\mathbf{R}_i$ will be $\{\lambda_{ij} - \kappa\}_{j=1}^{T_i}$. Then, choosing κ such that $\min\{\lambda_{ij}\} > \kappa > 0$ guarantees that Lemma 1 holds.

Because the previous decomposition is not unique, various values of κ will correspond to the same model and the same dynamics. Therefore, in practice the choice of κ will be based on convenience and numerical stability. One simple choice that has performed well in our simulations is to set $\kappa = \min\{\lambda_{ij}\}/2$. Also note that $\boldsymbol{\Omega}_i$ depends only on $\boldsymbol{\rho}$ and not on the data, so the decomposition need not be performed for every i.

Based on the preceding discussion, we present the estimation algorithm. For the sampling of the vector of AR coefficients $\boldsymbol{\rho} = (\rho_1, \ldots, \rho_p)'$, it is useful to define the following quantities. Let $e_{it} = z_{it} - \mathbf{x}_{it}'\boldsymbol{\beta} - \mathbf{w}_{it}'\mathbf{b}_i - g(s_{it})$, $\mathbf{e}_i = (e_{i,p+1}, \ldots, e_{i,T_i})'$, $\mathbf{e} = (\mathbf{e}_1', \ldots, \mathbf{e}_n')'$, and $\mathbf{E}$ denote the $(N - np) \times p$ matrix with rows containing p lags of e_{it} $(i = 1, \ldots, T_i,\ t \geq p + 1)$, that is, $(e_{i,t-1}, \ldots, e_{i,t-p})$. Finally, let the initial p values of e_{it} in each cluster be given by $\mathbf{e}_{i1} = (e_{i1}, \ldots, e_{ip})'$, and let $\boldsymbol{\Omega}_p$ be the $p \times p$ stationary covariance matrix of the AR(p) error process, which is a function of $\boldsymbol{\rho}$ and is constructed identically to $\{\boldsymbol{\Omega}_i\}$. Simulation of $\boldsymbol{\rho}$ is by the Metropolis–Hastings (M–H) algorithm (Hastings 1970; Tierney 1994; Chib and Greenberg 1995).

Algorithm 1: Model with state dependence and AR(p) *serial correlation.*

1. Sample $\{\mathbf{z}_i\}|\mathbf{y}, \mathbf{D}, \mathbf{g}, \boldsymbol{\beta}$, and $\boldsymbol{\rho}$ marginal of $\{\mathbf{b}_i\}$ by drawing, for $i \leq n$, $t \leq T_i$
$$z_{it} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{it}, v_{it}) & \text{if } y_{it} = 1 \\ \mathcal{TN}_{(-\infty,0]}(\mu_{it}, v_{it}) & \text{if } y_{it} = 0, \end{cases}$$
where $\mathcal{TN}_{(a,b)}(\mu_{it}, v_{it})$ is a normal distribution truncated to the interval (a, b) with mean $\mu_{it} = E(z_{it}|\mathbf{z}_{i\backslash t}, \boldsymbol{\beta}, \mathbf{g}_i, \mathbf{V}_i)$ and variance $v_{it} = \mathrm{var}(z_{it}|\mathbf{z}_{i\backslash t}, \boldsymbol{\beta}, \mathbf{g}_i, \mathbf{V}_i)$, with $\mathbf{V}_i = \boldsymbol{\Omega}_i + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$, and $\boldsymbol{\Omega}_i$ determined by $\boldsymbol{\rho}$ as discussed earlier.
2. Sample $\boldsymbol{\beta}, \{\mathbf{b}_i\}, \{\mathbf{u}_i\}|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g}_i$, and $\boldsymbol{\rho}$ in one block by drawing the following:
   (a) $\boldsymbol{\beta}|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g}, \boldsymbol{\rho} \sim \mathcal{N}(\hat{\boldsymbol{\beta}}, \mathbf{B})$, where $\hat{\boldsymbol{\beta}} = \mathbf{B}(\mathbf{B}_0^{-1}\boldsymbol{\beta}_0 + \sum_{i=1}^n \mathbf{X}_i'\mathbf{V}_i^{-1}(\mathbf{z}_i - \mathbf{g}_i))$ and $\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{X}_i'\mathbf{V}_i^{-1} \times \mathbf{X}_i)^{-1}$
   (b) $\mathbf{b}_i|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \boldsymbol{\beta}, \mathbf{g}, \boldsymbol{\rho} \sim \mathcal{N}(\hat{\mathbf{b}}_i, \mathbf{B}_i)$ with $\hat{\mathbf{b}}_i = \mathbf{B}_i\mathbf{W}_i' \times \boldsymbol{\Omega}_i^{-1}(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{g}_i)$ and $\mathbf{B}_i = (\mathbf{D}^{-1} + \mathbf{W}_i'\boldsymbol{\Omega}_i^{-1}\mathbf{W}_i)^{-1}$ for $i = 1, \ldots, n$
   (c) $\mathbf{u}_i|\mathbf{y}, \{z_{it}\}, \{\mathbf{b}_i\}, \boldsymbol{\beta}, \mathbf{g}, \boldsymbol{\rho} \sim \mathcal{N}(\hat{\mathbf{u}}_i, \mathbf{U}_i)$, where $\hat{\mathbf{u}}_i = \mathbf{U}_i\mathbf{C}_i(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{g}_i - \mathbf{W}_i\mathbf{b}_i)/\kappa$ and $\mathbf{U}_i = (\mathbf{I} + \mathbf{C}_i\mathbf{C}_i'/\kappa)^{-1}$ for $i = 1, \ldots, n$.
3. Sample $\mathbf{D}^{-1}|\{\mathbf{b}_i\} \sim \mathcal{W}_p\{r_0 + n, (\mathbf{R}_0^{-1} + \sum_{i=1}^n \mathbf{b}_i\mathbf{b}_i')^{-1}\}$.
4. Sample $\mathbf{g}|\mathbf{y}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \tau^2, \{z_{it}\}, \{\mathbf{u}_{i1}\} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$, where $\mathbf{G} = (\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{Q}/\kappa)^{-1}$ and $\hat{\mathbf{g}} = \mathbf{G}(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'(\mathbf{z} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\mathbf{b} - \mathbf{C}'\mathbf{u})/\kappa)$. Because $\mathbf{G}^{-1}$ is banded, estimation can proceed efficiently as discussed in Remark 1.
5. Sample $\tau^2|\mathbf{g} \sim \mathcal{IG}(\frac{\nu_0+m}{2}, \frac{\delta_0+(\mathbf{g}-\mathbf{g}_0)'\mathbf{K}(\mathbf{g}-\mathbf{g}_0)}{2})$.
6. Sample $\boldsymbol{\rho}|\mathbf{y}, \mathbf{g}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \{z_{it}\} \propto \Psi(\boldsymbol{\rho}) \times \mathcal{N}(\hat{\boldsymbol{\rho}}, \mathbf{P}) \times I_{S_\rho}$, where $\hat{\boldsymbol{\rho}} = \mathbf{P}(\mathbf{P}_0^{-1}\boldsymbol{\rho}_0 + \mathbf{E}'\mathbf{e})$, $\mathbf{P} = (\mathbf{P}_0^{-1} + \mathbf{E}'\mathbf{E})^{-1}$, and $\Psi(\boldsymbol{\rho}) = |\boldsymbol{\Omega}_p|^{-n/2}\exp(-\frac{1}{2}\sum_{i=1}^n \mathbf{e}_{i1}'\boldsymbol{\Omega}_p^{-1}\mathbf{e}_{i1})$.

In the M–H step of Algorithm 1, a proposal draw $\boldsymbol{\rho}'$ is generated from the density $\mathcal{N}(\hat{\boldsymbol{\rho}}, \mathbf{P})I_{S_\rho}$ and is subsequently accepted as the next sample value with probability $\min\{\Psi(\boldsymbol{\rho}')/\Psi(\boldsymbol{\rho}), 1\}$ (see Chib and Greenberg 1994). If the candidate value $\boldsymbol{\rho}'$ is rejected, then the current value $\boldsymbol{\rho}$ is repeated as the next value of the MCMC sample.

We note several aspects of this algorithm. First, because $\boldsymbol{\beta}$, $\{\mathbf{b}_i\}$, and $\{\mathbf{u}_i\}$ are correlated by construction, they are sampled in one block to speed up mixing of the chain. This is done by using (10) to sample $\boldsymbol{\beta}$ from a conditional density that does not depend on $\{\mathbf{b}_i\}$ and $\{\mathbf{u}_i\}$, followed by drawing $\{\mathbf{b}_i\}$ from a conditional density that depends on $\boldsymbol{\beta}$ but not on $\{\mathbf{u}_i\}$, and finally drawing $\{\mathbf{u}_i\}$ from its full-conditional density (Chib and Carlin 1999). Second, the MCMC approach to estimating τ^2 in this hierarchical model is an alternative to cross-validation that can be applied to both continuous and discrete data and fully accounts for parameter uncertainty, unlike plug-in approaches, which do not account for the variability due to estimating the smoothing parameter. An important alternative to the these methods is the maximum integrated likelihood approach to determining τ^2 (Kohn, Ansley, and Tharm 1991), but this is not feasible in this binary data setting, because of the N-dimensional integration that would be required.

Several extensions are possible. The approach for dealing with dependent errors can accommodate other correlation structures, such as exponentially correlated error sequences $\boldsymbol{\Omega}_i[t, s] = \exp\{-\alpha|t - s|^r\}$ for scalars α and r (e.g., Diggle and Hutchinson 1989). The method can also handle estimation of non-Toeplitz $\boldsymbol{\Omega}_i$ using the algorithms of Chib and Greenberg (1998). Yet other extensions can be pursued in the latent variable step; for example, the methods can be adapted to models for polychotomous data and models with t-links (Albert and Chib 1993) or to mixture-of-normals link functions, including an approximation to the logit link (Wood and Kohn 1998). Of course, Algorithm 1 subsumes an algorithm for the estimation of a simpler, fully parametric version of the model. Applications to continuous-data settings are also immediate.

4. AVERAGE COVARIATE EFFECTS

We now turn to the question of finding the effect of a change in a given covariate x_j. This is important for understanding the model and for determining the impact of an intervention on one or more of the covariates. However, a change in a covariate affects both the contemporaneous response and its future values. This effect also depends on all other covariates and model parameters. The impact is quite complex and nonlinear because of the nonlinear link function, the state dependence, the serial correlation, the unknown function, and the random effects. Hence it is calculated marginalized over the remaining covariates and the parameters.

Suppose that the model for a new individual i is given by

$$z_{it} = \mathbf{x}_{it}'\boldsymbol{\beta} + \mathbf{w}_{it}'\mathbf{b}_i + g(s_{it}) + \varepsilon_{it},$$

where the definitions of $\mathbf{x}_{it}$ and $\boldsymbol{\beta}$ are as in Sections 2 and 3, and we are interested in the effect of a particular x, say x_1, on contemporaneous and future y_{it}. Splitting $\mathbf{x}_{it}'$ and $\boldsymbol{\beta}$ accordingly, we rewrite the foregoing model as

$$z_{it} = x_{1it}\beta_1 + \mathbf{x}_{2it}'\boldsymbol{\beta}_2 + \mathbf{w}_{it}'\mathbf{b}_i + g(s_{it}) + \varepsilon_{it}.$$

The average covariate effect can then be analyzed from a predictive perspective applied to this new individual i. Given the specific context, we may consider various scenarios of interest; examples of economic policy interventions may include increasing or decreasing income by some percentage, and imposing an additional year of education. These interventions may affect the covariate values in a single period (e.g., a one-time tax break) or in multiple periods (e.g., a permanent tax reduction or increasing the minimum mandatory level of education). For specificity, suppose that one thinks of setting $\{x_{1it}\}_{t=1}^{T_i}$ to the values $\{x_{1it}^{\dagger}\}_{t=1}^{T_i}$. (Again, only a subset of these values could be affected by the intervention, whereas the others can remain unchanged.) For a predictive horizon of $t = 1, 2, \ldots, T_i$ (where T_i is the smallest of the cluster sizes in the observed data), one is now interested in the distribution of $y_{i1}, y_{i2}, \ldots, y_{iT_i}$ marginalized over $\{\mathbf{x}_{2it}\}$, $\mathbf{b}_i$, and $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\phi}, \mathbf{g}, \mathbf{D}, \tau^2, \boldsymbol{\rho})$ given the current data $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_n)$. A practical procedure is to marginalize out the covariates as a Monte Carlo average using their empirical distribution, while also integrating out $\boldsymbol{\theta}$ with respect to the posterior distribution $\pi(\boldsymbol{\theta}|\mathbf{y})$. Of course, $\mathbf{b}_i$ is independent of $\mathbf{y}$ and hence can be integrated out of the joint distribution of $\{z_{i1}, \ldots, z_{iT_i}\}$ analytically using the distribution $\mathcal{N}(\mathbf{0}, \mathbf{D})$, without recourse to Monte Carlo. Therefore, the goal is to obtain a sample of draws from the distribution

$$\begin{aligned}&\left[z_{i1}, \ldots, z_{iT_i}|\mathbf{y}, \{x_{1it}^{\dagger}\}\right] \\ &\quad = \int \left[z_{i1}, \ldots, z_{iT_i}|\mathbf{y}, \{x_{1it}^{\dagger}\}, \{\mathbf{x}_{2it}\}, \{\mathbf{w}_{it}\}, \{s_{it}\}, \boldsymbol{\theta}\right] \\ &\qquad \times \pi(\{\mathbf{x}_{2it}\}, \{\mathbf{w}_{it}\}, \{s_{it}\})\pi(\boldsymbol{\theta}|\mathbf{y})\, d\{\mathbf{x}_{it}\}\, d\{\mathbf{w}_{it}\}\, d\{s_{it}\}\, d\boldsymbol{\theta}.\end{aligned}$$

A sample from this predictive distribution can be obtained by the method of composition applied in the following way. Randomly draw an individual and extract the sequence of covariate values. Sample a value for $\boldsymbol{\theta}$ from the posterior density, and sample $\{z_{i1}, \ldots, z_{iT_i}\}$ jointly from $[z_{i1}, \ldots, z_{iT_i}|\mathbf{y}, \{x_{1it}^{\dagger}\}, \{\mathbf{x}_{2it}\}, \{\mathbf{w}_{it}\}, \{s_{it}\}, \boldsymbol{\theta}]$, constructing the $\{y_{it}\}$ in the usual way. Repeat this for other individuals and other draws from the posterior distribution to obtain the predictive probability mass function of $(y_{i1}, \ldots, y_{iT_i})$. Repeat this analysis for a different $\{x_{1it}\}$, say $\{x_{1it}^{\ddagger}\}$. The difference in the computed pointwise probabilities then gives the effect of x_1 as the values $\{x_{1it}^{\dagger}\}$ are changed to $\{x_{1it}^{\ddagger}\}$.

This approach can be similarly applied to other elements of $\mathbf{x}_{it}$ and $\mathbf{w}_{it}$. Quite significantly, it can be applied in determining the effect of the nonparametric component $g(s)$ because the error bands that are usually reported in the estimation of $g(\cdot)$ are pointwise, not joint, and do not provide sufficient information on which to make probabilistic statements about the functional shape (Hastie and Tibshirani 1990, p. 62), differences such as $g(s^{\dagger}) - g(s^{\ddagger})$, and the effect of s on the probability of response. In addition, we can condition on certain variables (gender, race, and specific initial conditions) that might determine a particular subsample of interest, in which case our procedures are applied only to observations in the subsample. An application of the techniques is considered in Section 7, where Figure 8 shows the estimated average covariate effects for husband's income and two child variables and Figure 9 shows these effects for two specific subsamples.

5. MODEL COMPARISON

A central issue in the analysis of statistical data is model formulation, because the appropriate specification is rarely known and is subject to uncertainty. Among other considerations, the uncertainty may be due to the problem of variable selection (i.e., the specific covariates and lags to be included in the model), the functional specification (a parametric model vs. a semiparametric model), or the distributional assumptions. In general, given the data $\mathbf{y}$, interest centers on a collection of models $\{\mathcal{M}_1, \ldots, \mathcal{M}_L\}$ representing competing hypotheses about $\mathbf{y}$. Each model $\mathcal{M}_l$ is characterized by a model-specific parameter vector $\boldsymbol{\theta}_l$ and sampling density $f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)$. Bayesian model selection proceeds by comparing the models in $\{\mathcal{M}_l\}$ through their posterior odds ratio, which for any two models $\mathcal{M}_i$ and $\mathcal{M}_j$ is written as

$$\frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \times \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)}, \tag{15}$$

where $m(\mathbf{y}|\mathcal{M}_l) = \int f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)\pi_l(\boldsymbol{\theta}_l|\mathcal{M}_l)\, d\boldsymbol{\theta}_l$ is the marginal likelihood of $\mathcal{M}_l$. The first fraction on the right side of (15) is known as the prior odds; the second is known as the Bayes factor.

To date, model comparisons in the semiparametric context have been based on such criteria as the AIC and BIC (e.g., Shively et al. 1999; Wood et al. 2002; DiMatteo et al. 2001; Hansen and Kooperberg 2002), because direct evaluation of the integral defining $m(\mathbf{y}|\mathcal{M}_l)$ is generally infeasible. But the AIC and BIC cannot be computed for the model in this article because of the difficulty (discussed in Sec. 3) of maximizing the likelihood in the semiparametric binary panel setting, whereas both the AIC and the BIC require the maximized value of the likelihood as input. Here we revisit the question of calculating the marginal likelihood of the semiparametric model and show that it can be managed through careful application of existing methods.

In particular, Chib (1995) provided a method based on the recognition that the marginal likelihood can be expressed as $m(\mathbf{y}|\mathcal{M}_l) = f(\mathbf{y}|\mathcal{M}_l, \boldsymbol{\theta}_l)\pi(\boldsymbol{\theta}_l|\mathcal{M}_l)/\pi(\boldsymbol{\theta}_l|\mathbf{y}, \mathcal{M}_l)$. This identity holds for any point $\boldsymbol{\theta}_l$, so that calculation of the marginal likelihood is reduced to finding an estimate of the posterior ordinate $\pi(\boldsymbol{\theta}_l^*|\mathbf{y}, \mathcal{M}_l)$ at a single point $\boldsymbol{\theta}_l^*$. In what follows, we suppress the model index for notational convenience. Suppose that the parameter vector $\boldsymbol{\theta}$ is split into B conveniently specified components or blocks (usually done on the basis of the natural groupings that emerge in constructing the MCMC sampler), so that

$\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_B)$. Let $\boldsymbol{\psi}_i^* = (\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_i^*)$ denote the blocks up to i, fixed at their values in $\boldsymbol{\theta}^*$, and let $\boldsymbol{\psi}^{i+1} = (\boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_B)$ denote the blocks beyond i. Then, by the law of total probability, we have

$$\pi(\boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_B^* | \mathbf{y}) = \prod_{i=1}^{B} \pi(\boldsymbol{\theta}_i^* | \mathbf{y}, \boldsymbol{\theta}_1^*, \ldots, \boldsymbol{\theta}_{i-1}^*)$$
$$= \prod_{i=1}^{B} \pi(\boldsymbol{\theta}_i^* | \mathbf{y}, \boldsymbol{\psi}_{i-1}^*).$$

When the full-conditional densities are known, each ordinate $\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*)$ can be estimated by Rao–Blackwellization as $\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*) \approx J^{-1} \sum_{j=1}^{J} \pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1,(j)})$, where $\boldsymbol{\psi}^{i,(j)} \sim \pi(\boldsymbol{\psi}^i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*), j = 1, \ldots, J$, come from a *reduced run* for $1 < i < B$, where sampling is only over $\boldsymbol{\psi}^i$, with the blocks $\boldsymbol{\psi}_{i-1}^*$ held fixed. The ordinate $\pi(\boldsymbol{\theta}_1^*|\mathbf{y})$ for the first block of parameters $\boldsymbol{\theta}_1$ is estimated with draws $\boldsymbol{\theta} \sim \pi(\boldsymbol{\theta}|\mathbf{y})$ from the main MCMC run, whereas the ordinate $\pi(\boldsymbol{\theta}_B^*|\mathbf{y}, \boldsymbol{\psi}_{B-1}^*)$ is available directly.

When one or more of the full conditional densities are not of a standard form and sampling requires the M–H algorithm, Chib and Jeliazkov (2001) used the local reversibility of the M–H chain to show that

$$\pi(\boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*) = \frac{E_1\{\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1}) q(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}}{E_2\{\alpha(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}}, \tag{16}$$

where E_1 is the expectation with respect to conditional posterior $\pi(\boldsymbol{\psi}^i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*)$ and E_2 is that with respect to the conditional product measure $\pi(\boldsymbol{\psi}^{i+1}|\mathbf{y}, \boldsymbol{\psi}_i^*) q(\boldsymbol{\theta}_i^*, \boldsymbol{\theta}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})$. In the preceding, $q(\boldsymbol{\theta}, \boldsymbol{\theta}'|\mathbf{y})$ denotes the candidate generating density of the M–H chain for moving from the current value $\boldsymbol{\theta}$ to a proposed value $\boldsymbol{\theta}'$, and $\alpha(\boldsymbol{\theta}_i, \boldsymbol{\theta}_i'|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})$ denotes the M–H probability of moving from $\boldsymbol{\theta}$ to $\boldsymbol{\theta}'$. Each of these expectations can be computed from the output of appropriately chosen reduced runs.

Three new and important considerations emerge when applying these methods to the model considered in this article. Our implementation, using Algorithm 1 to simulate the blocks $\{\mathbf{z}_i\}$, $\boldsymbol{\beta}$, $\{\mathbf{b}_i\}$, $\{\mathbf{u}_i\}$, $\mathbf{D}$, $\mathbf{g}$, τ^2, and $\boldsymbol{\rho}$, is based on the following posterior decomposition (marginalized over the latent data $\{\mathbf{z}_i\}$, $\{\mathbf{b}_i\}$, $\{\mathbf{u}_i\}$):

$$\pi(\mathbf{D}^*, \tau^{2*}|\mathbf{y}) \pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}) \times \pi(\boldsymbol{\rho}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*) \pi(\mathbf{g}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*).$$

The first consideration is that the ordinate of $\mathbf{g}$ is estimated last, because this tends to improve the efficiency of the ordinate estimation in the Rao–Blackwellization step, where

$$\pi(\mathbf{g}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*) \approx J^{-1} \sum_{j=1}^{J} \pi\big(\mathbf{g}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*, \{\mathbf{z}_i\}^{(j)}, \{\mathbf{b}_i\}^{(j)}, \{\mathbf{u}_i\}^{(j)}\big),$$

so that marginalization is only with respect to the conditional distribution of the latent data $\pi(\{\mathbf{z}_i\}, \{\mathbf{b}_i\}, \{\mathbf{u}_i\}|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*, \boldsymbol{\rho}^*)$, with all parameter blocks in the conditioning set fixed. Because the simulation algorithm for $\mathbf{g}$ is $\mathcal{O}(m)$, this particular choice comes at a manageable computational cost (from having to simulate $\mathbf{g}$ in all of the preceding reduced runs), whereas the statistical efficiency benefits may be substantial when m is large. Second, it should also be noted that the ordinate $\pi(\mathbf{D}^*, \tau^{2*}|\mathbf{y})$ can be estimated jointly because, conditional on $\{\mathbf{b}_i\}$ and $\mathbf{g}$, the full conditional densities of $\mathbf{D}$ and τ^2 are independent. This observation saves a reduced run. Third, in Algorithm 1 the proposal density $q(\boldsymbol{\rho}, \boldsymbol{\rho}'|\mathbf{y}, \cdot) = q(\boldsymbol{\rho}'|\mathbf{y}, \cdot)$ is a truncated normal density with an unknown normalizing constant except in the AR(1) case. Specifically, over the region of stationarity $S_{\boldsymbol{\rho}}$, we have $q(\boldsymbol{\rho}|\mathbf{y}, \cdot) = h(\boldsymbol{\rho}|\mathbf{y}, \cdot) / \int_{S_{\boldsymbol{\rho}}} h(\boldsymbol{\rho}|\mathbf{y}, \cdot)\, d\boldsymbol{\rho}$, where $h(\boldsymbol{\rho}|\mathbf{y}, \cdot)$ is an unrestricted normal density for $\boldsymbol{\rho}$. We avoid the need to compute this unknown constant of integration by noting that when the reversibility condition used by Chib and Jeliazkov (2001) to obtain (16) is written in terms of $q(\boldsymbol{\rho}|\mathbf{y}, \cdot)$, its unknown normalizing constant (being the same on both sides) will cancel out, so that, on integration, we have

$$\pi(\boldsymbol{\rho}^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*) = \frac{E_1\{\alpha(\boldsymbol{\rho}, \boldsymbol{\rho}^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1}) h(\boldsymbol{\rho}^*|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}}{E_2\{\alpha(\rho_i^*, \boldsymbol{\rho}_i|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})\}},$$

where $\boldsymbol{\psi}_{i-1}^* = (\mathbf{D}^*, \tau^{2*}, \boldsymbol{\beta}^*)$, $\boldsymbol{\psi}^{i+1} = (\mathbf{g}, \{\mathbf{z}_i\}, \{\mathbf{b}_i\}, \{\mathbf{u}_i\})$, E_1 is the expectation with respect to $\pi(\boldsymbol{\rho}, \boldsymbol{\psi}^{i+1}|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*)$, and E_2 is the expectation with respect to $\pi(\boldsymbol{\psi}^{i+1}|\mathbf{y}, \boldsymbol{\psi}_i^*) h(\boldsymbol{\rho}|\mathbf{y}, \boldsymbol{\psi}_{i-1}^*, \boldsymbol{\psi}^{i+1})$. It is also useful to note that because $h(\boldsymbol{\rho}'|\mathbf{y}, \cdot)$ does not depend on the current value of $\boldsymbol{\rho}$ in the sampler, estimating the denominator quantity is done with draws available from the same run in which the numerator is estimated.

Implementing these methods requires the likelihood ordinate $f(\mathbf{y}|\mathbf{D}^*, \boldsymbol{\beta}^*, \mathbf{g}^*, \boldsymbol{\rho}^*)$, which we obtain by the method of Geweke, Hajivassiliou, and Keane (GHK) (Geweke 1991; Börsch-Supan and Hajivassiliou 1993; Keane 1994), using 10,000 Monte Carlo iterations.

6. SIMULATION STUDY

We carried out a simulation study to examine the performance of the techniques proposed in Algorithm 1. For the estimates of the nonparametric function, we report mean squared errors (MSEs) for several designs. The posterior mean estimates $E\{g(\mathbf{v})|\mathbf{y}\}$, are found from MCMC runs of length 5,000 after burn-ins of 1,000 cycles. For the parametric components of the model, we report the autocorrelations and inefficiency factors under alternative specifications and sample sizes. We find that the overall performance of the MCMC algorithm improves with larger sample sizes (increasing either the number of clusters n or the cluster sizes $\{T_i\}$), and that the random effects are simulated better when the increase in sample size comes from larger $\{T_i\}$.

Data are simulated from the model in (1) and (7), without serial correlation, using one, two, and three state-dependence lags, a single fixed-effect covariate $\tilde{\mathbf{X}}$, and one or two individual-effect covariates $\mathbf{W}$ (including a random intercept) that are correlated with (the average of) the initial conditions. $\tilde{\mathbf{X}}$ and $\mathbf{W}$ contain independent standard normal random variables, and we use $\boldsymbol{\delta} = 1$, $\boldsymbol{\gamma} = \mathbf{1}$, $\boldsymbol{\phi} = .5 \times \mathbf{1}$, and $\mathbf{D} = .2 \times \mathbf{I}$. We generate panels with 250, 500, and 1,000 clusters and with 10 time periods, using only the last 7 for estimation ($T_i = 7$, $i = 1, \ldots, n$), because our largest models contain 3 lags. We consider three functional forms for $g(s)$, presented in Figure 1, which capture a range of possible specifications in the literature. For the simulations, each function is evaluated on a regular grid of $m = 51$ points.

We gauge the performance of the method in fitting these functions using $MSE = \frac{1}{m}\sum_{j=1}^{m}\{\hat{g}(v_j) - g(v_j)\}^2$. The average MSE, together with the standard errors based on 20 data samples, is reported in Table 1 for alternative designs and sample sizes. In all cases, we have used comparable mild priors $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 10\mathbf{I})$, $(g_1, g_2)'|\tau^2 \sim \mathcal{N}(\mathbf{0}, \tau^2 100 \times \mathbf{I})$, $\mathbf{D}^{-1} \sim \mathcal{W}(q + 4, 1.67 \times \mathbf{I}_q)$, and $\tau^2 \sim \mathcal{IG}(3, .1)$; the priors on the variance parameters imply that $E(\mathbf{D}) = .2 \times \mathbf{I}$, $\mathrm{SD}(\mathrm{diag}(\mathbf{D})) = .283 \times \mathbf{1}$, $E(\tau^2) = .05$, $\mathrm{SD}(\tau^2) = .05$, and $E_{\tau^2}(\mathrm{var}((g_1, g_2)')) = 5 \times \mathbf{I}$. (We discuss the role of the smoothness prior in more detail later.) Table 1 also shows that as the sample size grows, the functions are estimated more precisely, as expected. Also, in line with conventional wisdom, the general trend seems to be that fitting models with fewer parameters results in lower MSE estimates. We clarify that under this setup, increasing the number of lags J affects the simulation study in two ways: first, the dimension of the parameter space increases, and second, it affects the proportion of 1's among the responses (because all elements in $\boldsymbol{\phi}$ are positive). It is well known that the degree of asymmetry in the proportion of the responses affects the estimation precision. The proportion of 1's is between .62 and .67 across the three functional specifications for our one-lag models, between .67 and .72 for the two-lag models, and between .71 and .76 for the three-lag models. As Table 1 shows, however, the method recovers the true functions well despite this asymmetry. The computational cost per 1,000 MCMC draws is approximately 16 seconds in the simplest case ($n = 250$, $J = 1$, $q = 1$). Adding more fixed effects (e.g., lags) does not increase the costs by more than 1/10 of a second. However, adding an additional random effect increases the computational cost to 58 seconds per 1,000 draws, because of the expense of simulating a larger random-effects vector in each cluster. Finally, the computational times increased linearly in relation to the sample size n. Figure 2 presents three particular function fits for the case ($n = 500$, $J = 2$, $q = 1$).

Table 1. Average Mean Squared Errors (with estimated standard errors in parentheses)

Clusters	Lags	Random effects	Average MSE ($\times 10^{-2}$)		
			g_1	g_2	g_3
$n = 250$	$J = 1$	$q = 1$	$1.60_{(.24)}$	$1.40_{(.14)}$	$1.77_{(.32)}$
		$q = 2$	$1.92_{(.26)}$	$1.85_{(.23)}$	$1.41_{(.19)}$
	$J = 2$	$q = 1$	$2.21_{(.43)}$	$1.61_{(.21)}$	$1.78_{(.37)}$
		$q = 2$	$2.72_{(.48)}$	$2.82_{(.40)}$	$2.62_{(.48)}$
	$J = 3$	$q = 1$	$2.89_{(.99)}$	$1.98_{(.54)}$	$2.93_{(.41)}$
		$q = 2$	$3.42_{(.67)}$	$2.12_{(.60)}$	$3.02_{(.75)}$
$n = 500$	$J = 1$	$q = 1$	$.68_{(.08)}$	$.88_{(.14)}$	$1.04_{(.21)}$
		$q = 2$	$.98_{(.15)}$	$1.16_{(.13)}$	$.98_{(.09)}$
	$J = 2$	$q = 1$	$.96_{(.15)}$	$.91_{(.23)}$	$1.05_{(.18)}$
		$q = 2$	$1.43_{(.17)}$	$1.69_{(.28)}$	$1.57_{(.25)}$
	$J = 3$	$q = 1$	$2.57_{(.41)}$	$1.65_{(.38)}$	$1.81_{(.26)}$
		$q = 2$	$1.82_{(.31)}$	$2.49_{(.58)}$	$1.76_{(.22)}$
$n = 1{,}000$	$J = 1$	$q = 1$	$.64_{(.12)}$	$.63_{(.11)}$	$.60_{(.10)}$
		$q = 2$	$.56_{(.07)}$	$.58_{(.08)}$	$.68_{(.10)}$
	$J = 2$	$q = 1$	$.83_{(.13)}$	$.52_{(.09)}$	$.72_{(.11)}$
		$q = 2$	$.71_{(.11)}$	$.69_{(.17)}$	$.57_{(.08)}$
	$J = 3$	$q = 1$	$1.27_{(.29)}$	$.60_{(.09)}$	$.65_{(.11)}$
		$q = 2$	$.94_{(.19)}$	$1.19_{(.28)}$	$.84_{(.12)}$

NOTE: The three functions used in generating data for the simulation study were $g_1(s) = \sin(2\pi s)$, $s \in [.6, 1.4]$; $g_2(s) = -1 + s + 1.6s^2 + \sin(5s)$, $s \in [0, 1.1]$; $g_3(s) = -.8 + s + \exp\{-30(s - .5)^2\}$, $s \in [0, 1]$.

We now comment on the influence of the prior on our function estimates. The model as a whole has three important variance structures, and the relative informativeness of the priors for each of these structures should be viewed in the context of the other two structures, not in isolation. As already discussed, the variance of the errors and the variance τ^2 of the Markov process prior determine the trade-off between a good fit and a smooth function $g(\cdot)$. Similarly, the variance of the errors and

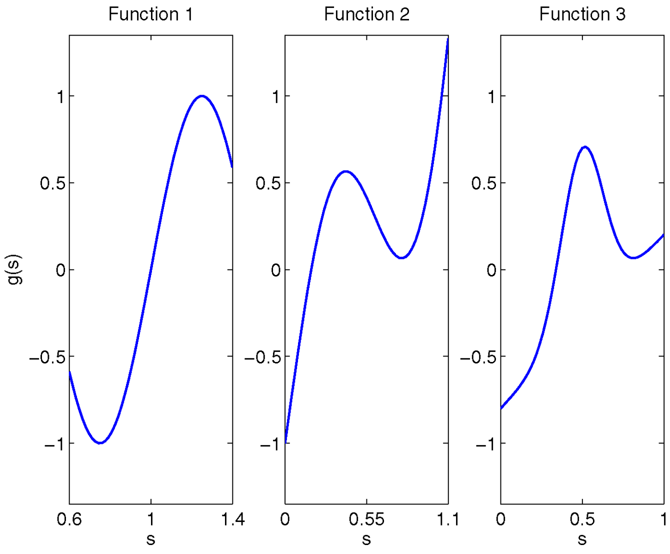

Figure 1. The Three Functions Used in Generating Data for the Simulation Study: g(s) = sin(2πs), s ∈ [.6, 1.4]; g(s) = −1 + s + 1.6s² + sin(5s), s ∈ [0, 1.1]; g(s) = −.8 + s + exp{−30(s − .5)²}, s ∈ [0, 1].

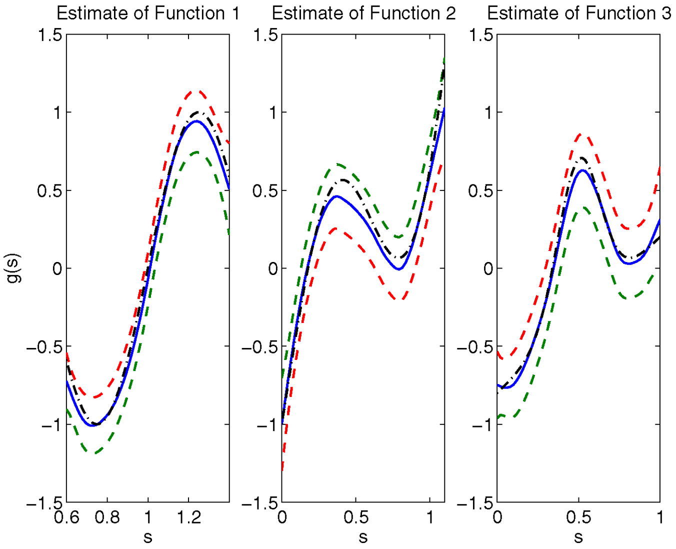

Figure 2. Simulation Study. Three examples of estimated functions (——), true functions (-·-·-), and 95% pointwise confidence bands (- - - -).

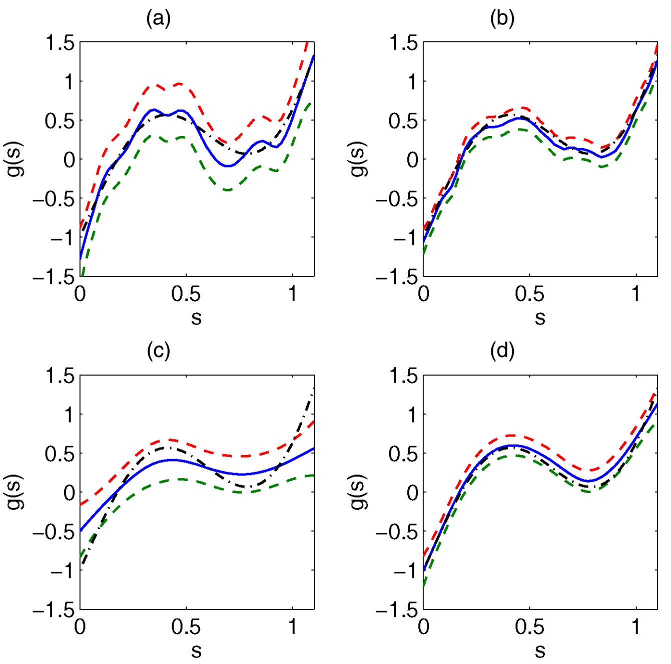

Figure 3. Effect of τ^2 on the Estimates of **g**: *$n = 250$ [(a) and (c)] and $n = 1{,}000$ [(b) and (d)]. A larger τ^2 is used in (a) and (b) than in (c) and (d). The dashed lines represent pointwise confidence bands.*

the variance of the random effects **D** determine a balance between intracluster and intercluster variation. Whereas in large samples the effects of the assumed priors on the parameter estimates is small (vanishing asymptotically), informative priors do matter in small samples. Figure 3 illustrates this point by using two exaggeratedly different informative priors on τ^2. In one case the prior on τ^2 is such that $E(\tau^2) = .5$ and $\mathrm{SD}(\tau^2) = .1$; in the second case $E(\tau^2) = .001$ and $\mathrm{SD}(\tau^2) = .001$. Figure 3 illustrates that the first prior leads to a function that becomes more wiggly as it curves to interpolate the data more closely, whereas the second prior leads to oversmoothing. When the sample size is increased from $n = 250$ to $n = 1{,}000$, the difference in the function estimates becomes smaller.

Next we consider the performance of the MCMC algorithm in fitting the parametric part of the model. For example, the case where $n = 500$ (with $T_i = 7$) is illustrated in Figure 4, which shows histograms and kernel-smoothed marginal posterior densities of the parameters together with the corresponding autocorrelations from the sampled output. The linear effects, together with τ^2, appear to be estimated well and the output is characterized by low autocorrelations. Although it can be seen that **D** is estimated well, its higher autocorrelation indicates slower mixing than that of the remaining parameters, so that longer MCMC runs may be needed to accurately describe the marginal posterior density of **D**. The slower mixing occurs because **D** is a parameter at the second level of the modeling hierarchy and depends on the data only indirectly through $\{\mathbf{b}_i\}$. Because the $\{\mathbf{b}_i\}$ are not well identified in smaller clusters, when only a few observations are available to identify the cluster-specific effects, and because learning about **D** occurs from the intercluster variation of $\{\mathbf{b}_i\}$, **D** also suffers from weak identification when cluster sizes are small. To measure the efficiency of the MCMC parameter sampling scheme, we use the measures $[1 + 2\sum_{k=1}^{L}\rho_k(l)]$, where $\rho_k(l)$ is the sample

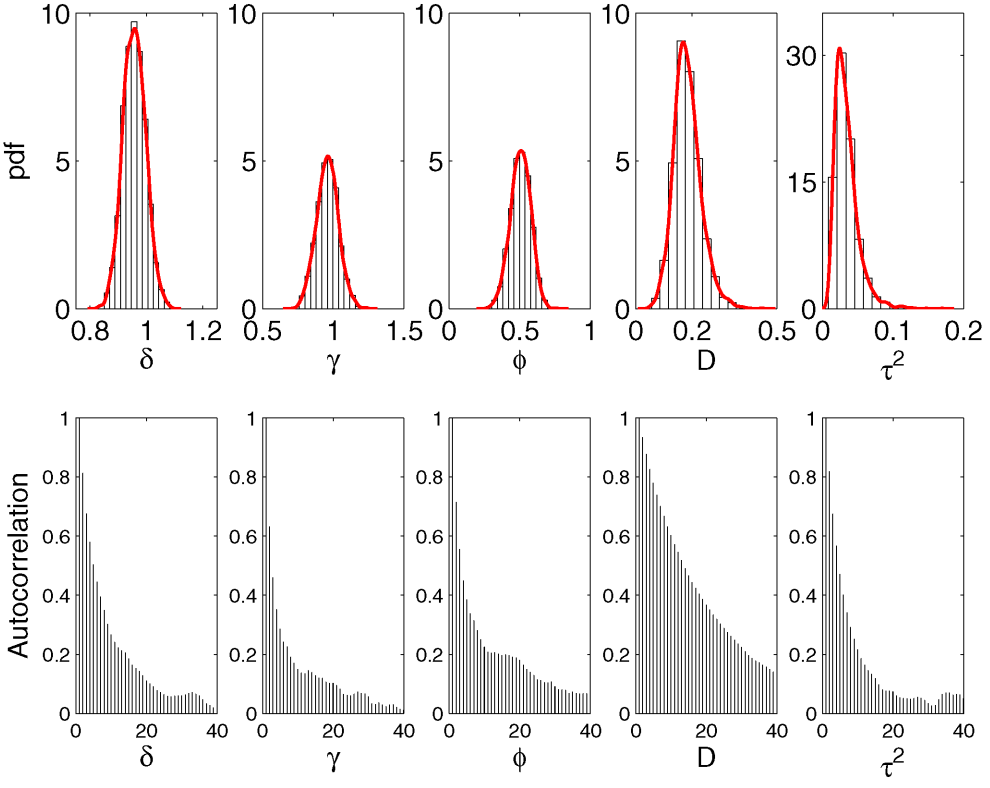

Figure 4. Posterior Samples and Autocorrelations for the Parameters of a Semiparametric Model With One Fixed Effect, One Random Effect, and One Lag ($T_i = 7$, $i = 1, \ldots, n$).

Table 2. Examples of Estimated Inefficiency Factors (autocorrelation times) for the Parameters of a Model With One Lag, One Random Effect, and One Fixed Effect for n = 500

	Inefficiency factor		
Parameter	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	13.68	10.18	12.55
γ	9.62	6.16	7.25
ϕ	12.11	8.31	9.03
$\mathbf{D}$	24.00	13.28	9.41
τ^2	10.65	5.26	8.74

autocorrelation at lag l for the kth parameter in the sampling with the summation truncated at values L at which the correlations taper off. The latter quantity, called the *inefficiency factor*, 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. Table 2 gives the inefficiency factors corresponding to the parameters for the same model as before, but now with different cluster sizes (7, 12, and 17 observations per cluster). In this setup the larger cluster sizes serve to identify $\{\mathbf{b}_i\}$ better, allowing more precise capture of intercluster variation. Table 2 shows that the inefficiency factor for **D** drops considerably (but the other inefficiency factors stay within a similar range). The improved sampling of **D** becomes evident from comparing Figures 4 and 5, with the latter summarizing the MCMC output when $T_i = 17$.

Similar to the cases given in Table 2, Table 3 presents results for the inefficiency factors for a model with two lags and two random effects. Because now not one, but two, random effects are estimated from the limited observations in each cluster, the elements of **D** are sampled with somewhat higher inefficiency factors. Here again, however, Table 3 shows that as the cluster sizes increase, resulting in better identification of $\{\mathbf{b}_i\}$, the inefficiency factors for the elements of the heterogeneity matrix **D** drop noticeably.

In the rest of this section, we report results from experiments involving a model with AR(1) correlated errors. Table 4 presents the inefficiency factors for three cluster sizes and two values of ρ. The parameters ρ and **D** are sampled well in all cases, but it is interesting to see that when ρ is positive, both ρ and **D** have higher inefficiency factors than otherwise. This is because both are estimated from the covariance of the errors, and decomposing that matrix into an equicorrelated part (with positive elements implied by the random intercept) and a Toeplitz part [implied by the AR(1) part, which also has positive elements when $\rho > 0$] is difficult in small samples. As the cluster sizes increase, **D** is identified better, so both ρ and **D** are estimated better. This does not appear to be a problem when $\rho < 0$, because then the two correlation structures implied by ρ and **D** are quite different. For the samplers and values of ρ considered here, the M–H acceptance rate in the sampling of ρ is in the range of (.87, .98).

These results show how serial correlation in the errors affects the performance of the sampler, but it is also of interest to note that misspecifying the correlation structure has definite impact on the remaining components of the model. As mentioned earlier, several studies, including those of Diggle and Hutchinson (1989), Altman (1990), and Smith et al. (1998), have pointed out that serial correlation, if incorrectly ignored, can have substantial adverse consequences for the estimation of the nonparametric function. What is interesting in the context of panel data

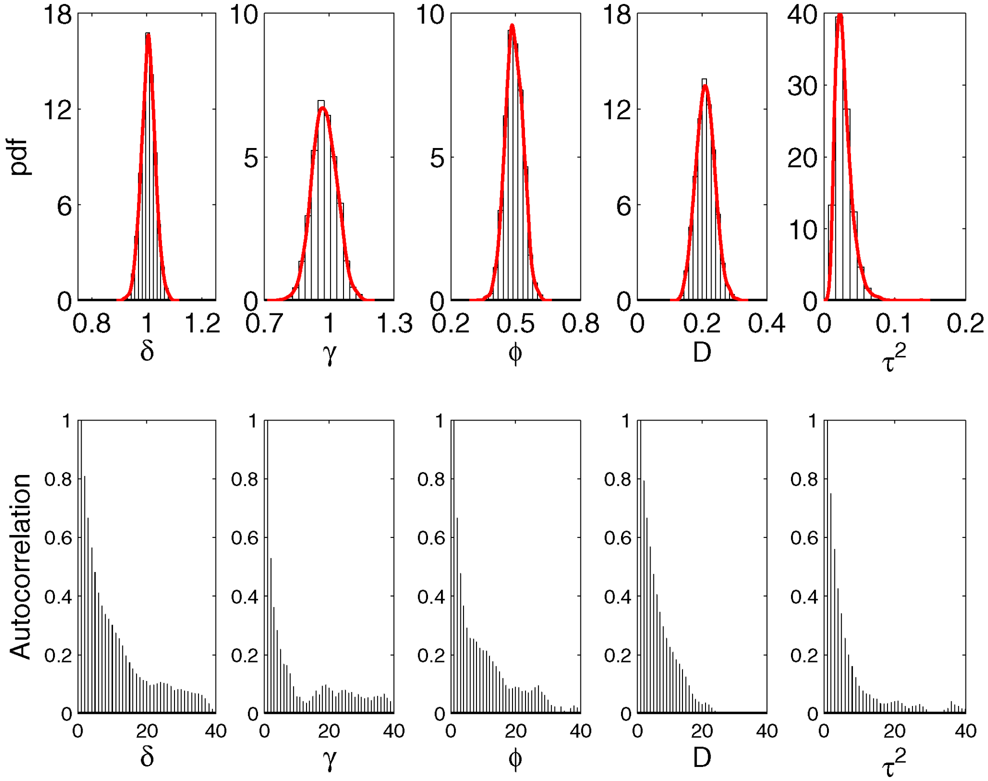

Figure 5. Posterior Samples and Autocorrelations for the Parameters of a Semiparametric Model With One Fixed Effect, One Random Effect, and One Lag ($T_i = 17$, $i = 1, \ldots, n$).

Table 3. Examples of Estimated Inefficiency Factors for the Parameters of a Model With Two Lags, Two Random Effects, and One Fixed Effect

	Inefficiency factor								
	n = 250			*n = 500*			*n = 1,000*		
Parameter	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	21.16	18.09	18.88	19.89	16.07	14.72	20.79	22.46	20.65
γ_1	12.19	12.20	11.51	14.73	13.00	12.82	12.16	13.57	12.34
γ_2	11.40	8.48	9.61	15.64	8.84	7.33	14.64	10.07	9.56
γ_3	9.06	10.36	8.41	11.75	10.48	9.19	12.59	12.90	8.44
ϕ_1	9.29	9.12	8.00	7.01	10.96	11.07	14.36	10.23	10.81
ϕ_2	8.39	7.60	10.14	7.69	7.52	8.48	9.15	11.14	10.01
$\mathbf{D}_{11}$	27.01	22.46	18.87	31.56	23.88	18.72	34.10	25.95	18.08
$\mathbf{D}_{12}$	29.73	23.16	18.16	29.88	25.81	14.91	26.31	20.25	18.38
$\mathbf{D}_{22}$	26.24	26.37	17.63	34.72	27.87	18.09	34.56	24.67	21.80
τ^2	10.58	12.45	10.17	14.74	9.47	7.80	8.02	6.05	6.02

is that ignoring the serial correlation can lead to biases in the estimates of the heterogeneity matrix **D**. In line with the discussion in the foregoing paragraph, we have found that ignoring the serial correlation distorts the estimates of **D**, especially with smaller cluster sizes or positive and high serial correlation. Ignoring the serial correlation in the errors also produces estimates of the lag coefficients $\boldsymbol{\phi}$ that differ widely from the true values used in generating the data (and in some settings having the opposite sign). One should also keep in mind that differences also can occur because of the usual identification restriction in binary data models, namely that the error variance is fixed. For example, if the errors follow the AR(1) process $\varepsilon_{it} = \rho\varepsilon_{i,t-1} + \nu_{it}$ with $\text{var}(\nu_{it}) = 1$, then the unconditional variance of ε_{it} is inflated to $1/(1-\rho^2)$. This is one of the points raised by Smith et al. (1998) in the context of a continuous-data model, where this larger error variance was shown to produce less efficient nonparametric function estimates. In the context of a binary data problem, ignoring serial correlation and setting $\text{var}(\varepsilon_{it}) = 1$ then corresponds to implicitly rescaling the regression parameters by a factor of $\sqrt{1-\rho^2}$. We take all of these results together as a strong signal not to ignore the modeling of the three causes of intertemporal dependence discussed in Section 1—state dependence due to lags of the dependent variable, serial correlation in the errors, and heterogeneity among clusters.

In terms of computational intensity, our method was quite efficient. For example, in one of our smaller datasets ($n = 250$, $q = 1$, $J = 1$, and $T_i = 7$, $i = 1, \dots, n$), Algorithm 1 took approximately 22 seconds to produce 1,000 MCMC draws, varying by less than .2 second as we increased the dimension of $\mathbf{g}$ from $m = 50$ to $m = 200$. For comparison, we also tried a "brute-force" algorithm that did not exploit banded matrices or our approach for dealing with correlated errors. In this case the computational cost of 1,000 MCMC draws was 35 seconds for $m = 50$, 103 seconds for $m = 100$, and 405 seconds for $m = 200$. We suspect that the brute-force algorithm would become largely infeasible in even higher-dimensional problems because of its computational intensity and storage requirements.

Table 4. Examples of Estimated Inefficiency Factors for the Parameters of the Model With One Lag, One Random Effect, One Fixed Effect, and AR(1) Serial Correlation for n = 500

	Inefficiency factor ($\rho = -.5$)			*Inefficiency factor ($\rho = .5$)*		
Parameter	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	19.06	17.43	20.19	20.75	14.06	17.72
γ	9.34	7.30	7.86	11.42	11.30	12.05
ϕ	23.05	25.63	18.24	17.00	20.24	20.11
D	21.48	16.21	12.75	39.47	28.02	15.19
τ^2	9.68	7.08	6.01	11.84	7.97	7.97
ρ	20.50	21.35	20.28	32.40	27.12	26.45

To summarize, the results suggest that the MCMC algorithm performs well, and that the estimation method recovers the parameters and functions used to generate the data. The performance of the method in recovering the nonparametric function $g(\cdot)$ and the model parameters improves with the sample size, when the model is better identified. Most noticeably, the sampling of **D** benefits strongly from larger cluster sizes.

7. INTERTEMPORAL LABOR FORCE PARTICIPATION OF MARRIED WOMEN

In this section we consider an application to the annual labor force participation decisions of 1,545 married women in the age range 17–66. The dataset, obtained from the PSID, is based on the work of Hyslop (1999) and contains a panel of women's working status indicators (1, working during the year; 0, not working) over a 7-year period (1979–1985), together with a set of seven covariates given in Table 5. The sample consists of continuously married couples where the husband is a labor force participant (reporting both positive earnings and hours worked) in each of the sample years. Similar data have been analyzed by Chib and Greenberg (1998), Avery, Hansen, and Hotz (1983), and Hyslop (1999) using other models and estimation techniques.

A key feature of our modeling is that the effect of age on the conditional probability of working is specified nonparametrically. There are compelling reasons for doing this. Nonlinearities arise because of changes in tastes, trade-offs, and age-related health over a woman's life cycle; due to the fact that age is indicative of the expected timing of events (e.g., graduation from school or college, planning for children); and because age may be revealing of social values and education type (cohort effect) or experience as a homemaker and in the labor market (productivity effect).

To account for model uncertainty, we evaluated several competing models that differ in their state dependence, serial correlation, and heterogeneity using the following baseline

Table 5. Variables in the Women's Labor Force Participation Application

Variable	*Explanation*	*Mean*	*SD*
WORK	Wife's labor force status (1, working; 0, not working)	.7097	.4539
INT	An intercept term (a column of 1's)		
AGE	The woman's age in years	36.0262	9.7737
RACE	1 if black, 0 otherwise	.1974	.3981
EDU	Attained education (in years) at time of survey	12.4858	2.1105
CH2	Number of children age 0–2 in that year	.2655	.4981
CH5	Number of children age 3–5 in that year	.3120	.5329
INC	Total annual labor income of head of household	31.7931	22.6417

NOTE: INC (in thousands of dollars) is measured as nominal earnings adjusted by the Consumer Price Index (base year 1987).

specification:

$$\begin{aligned} y_{it} &= \mathbb{1}\{\tilde{\mathbf{x}}_{it}'\boldsymbol{\delta} + \mathbf{w}_{it}'\boldsymbol{\beta}_i + g(s_{it}) \\ &\qquad + \phi_1 y_{i,t-1} + \cdots + \phi_J y_{i,t-J} + \varepsilon_{it} > 0\}, \\ \boldsymbol{\beta}_i &= \mathbf{A}_i\boldsymbol{\gamma} + \mathbf{b}_i, \qquad \mathbf{b}_i \sim \mathrm{N}_3(\mathbf{0}, \mathbf{D}), \end{aligned}$$

where $y_{it} = \mathrm{WORK}_{it}$, $\tilde{\mathbf{x}}_{it}' = (\mathrm{RACE}_i, \mathrm{EDU}_{it}, \ln(\mathrm{INC}_{it}))$, $s_{it} = \mathrm{AGE}_{it}$, $\mathbf{w}_{it}' = (1, \mathrm{CH2}_{it}, \mathrm{CH5}_{it})$, ε_{it} are potentially serially correlated, and the effects of CH2 and CH5 are allowed to depend on husbands' earnings and the initial conditions through

$$\mathbf{A}_i = \begin{pmatrix} \bar{y}_{i0} & & & & & & \\ & 1 & \bar{y}_{i0} & \overline{\ln(\mathrm{INC}_i)} & & & \\ & & & & 1 & \bar{y}_{i0} & \overline{\ln(\mathrm{INC}_i)} \end{pmatrix}.$$

Such issues as variable selection, lag determination, and correlation between the unobserved effects and covariates are handled as model selection problems by computing the marginal likelihoods of competing models. The semiparametric models are also compared against two parametric alternatives. The more important competing models are presented in Table 6.

Models $\mathcal{M}_1$–$\mathcal{M}_4$ have differing dynamics. Table 6 shows that allowing for AR(1) errors improves the performance of the one-lag state dependence model; serial correlation is present in the errors of that model (with ρ estimated as $-.288$, with a posterior standard deviation of .048). However, the data in this application strongly favor a model in which state dependence is incorporated through two lags of the dependent variable. Single-lag models ($\mathcal{M}_1$ and $\mathcal{M}_2$) have marginal likelihoods much lower than the corresponding two-lag versions ($\mathcal{M}_3$ and $\mathcal{M}_4$). We also see that inclusion of the second lag removes the serial correlation in the errors and results in a marginal likelihood of $\mathcal{M}_3$ that is highest among the models considered in Table 6. For model $\mathcal{M}_4$, the estimated value of ρ is $-.047$ with a 95% confidence interval $(-.224, .113)$. Models with higher-order dynamic dependence were considered, but received less support than $\mathcal{M}_3$. Regarding the heterogeneity in the model, we see from Table 6 that the single unobserved effect model ($\mathcal{M}_5$) is decisively less supported than $\mathcal{M}_3$.

We now discuss models $\mathcal{M}_5$ and $\mathcal{M}_6$ in the context of the estimate of the nonparametric function $g(\mathrm{AGE})$ from model $\mathcal{M}_3$. Figure 6 shows that the estimated effect of age departs notably from linearity. To examine whether a parametric model can adequately fit the data, we consider two parametric models ($\mathcal{M}_5$ and $\mathcal{M}_6$), where $g(\mathrm{AGE})$ is restricted to be linear or quadratic. [Because now $g(\cdot)$ is not general, $\mathbf{A}_i$ is not restricted for identification purposes.] The comparisons are shown in Figure 7. The estimates suggest that the linear model can be deceiving, because it produced a negative coefficient estimate for age of $-.0105$ with 95% credibility region given by $(-.018, -.003)$. The quadratic model comes closer to the semiparametric fit, but still leaves some excess nonlinearity undetected. Both parametric models miss the large increase in the probability of working in the early twenties and the nonlinearity around age 30. The marginal likelihood of the parametric

Table 6. Log Marginal Likelihoods for Alternative Models in the Women's Labor Force Participation Application (estimated from MCMC runs of length 15,000)

Model	*Fixed effect*	*Random effect*	*Nonzero elements in* $\mathbf{A}_i$	*ln(marginal likelihood)*
Model with 1*-lag state dependence and independent errors*:				
$\mathcal{M}_1$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}$	$\mathbf{w}_{it}$	$(\bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{\ln(\mathrm{INC}_i)};$ $1, \bar{y}_{i0}, \overline{\ln(\mathrm{INC}_i)})$	−2,610.7
Model with 1*-lag state dependence and* AR(1) *errors*:				
$\mathcal{M}_2$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	−2,595.5
Model with 2*-lag state dependence and independent errors*:				
$\mathcal{M}_3$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	−2,563.8
Model with 2*-lag state dependence and* AR(1) *errors*:				
$\mathcal{M}_4$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	−2,579.6
Model with random intercept only:				
$\mathcal{M}_5$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}, y_{i,t-2}$	1	$(\bar{y}_{i0})$	−2,580.1
Parametric versions of $\mathcal{M}_3$:				
$\mathcal{M}_6$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}, y_{i,t-2},$ AGE_{it}	$\mathbf{w}_{it}$	$(1, \bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{\ln(\mathrm{INC}_i)};$ $1, \bar{y}_{i0}, \overline{\ln(\mathrm{INC}_i)})$	−2,581.2
$\mathcal{M}_7$	$\tilde{\mathbf{x}}_{it}, y_{i,t-1}, y_{i,t-2},$ $\mathrm{AGE}_{it}, \mathrm{AGE}_{it}^2$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_6$	−2,574.6

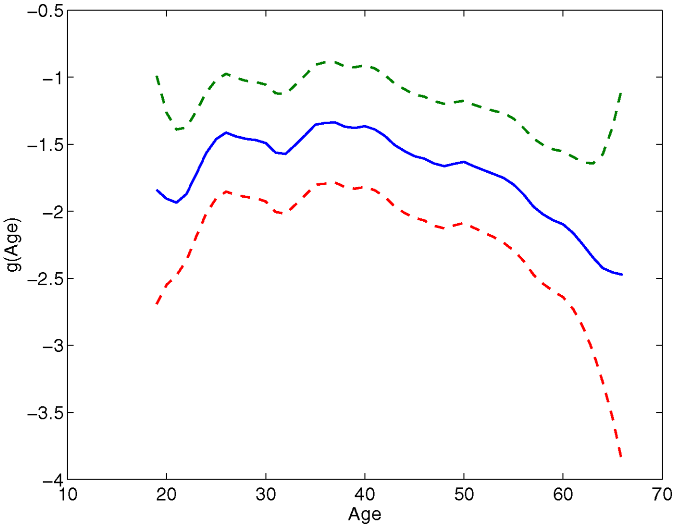

Figure 6. The Effect of Age on the Probability of Working: Nonparametric Estimate and Pointwise Confidence Bands (— E{g(AGE)|y}; - - upper; - - lower).

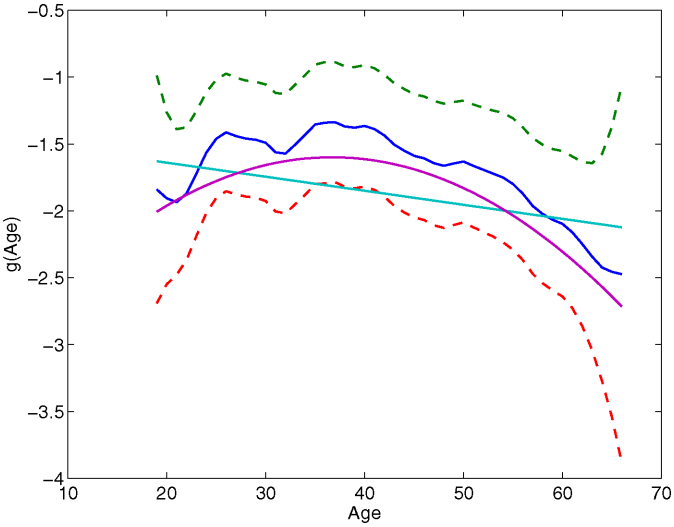

Figure 7. Comparison of the Linear (—), Quadratic (—), and Nonparametric (—) Estimates of g(AGE), With Pointwise Confidence Bands (- -).

models, especially that of the linear model, are lower confirming the conjecture from Figure 7 that these models do not fully account for the nonlinear effect of age.

Table 7 contains the parameter estimates of the best-fitting model ($\mathcal{M}_3$). We see that, conditional on the covariates, black women, better-educated women, and women whose husbands have low earnings are more likely to work. The results also indicate the strong state dependence on two lags of a woman's employment status. After controlling for state dependence and the remaining covariates, we see that the presence of children has a larger effect on the probability of working when the husband's earnings increase. Finally, the positive correlation between the random intercept and the initial conditions is consistent with the notion that the initial observations are indicative of a woman's tastes and human capital. The inefficiency factors in Table 7 (defined in Sec. 6) indicate a good overall performance of the MCMC sampler. But because of the large number of random effects and small cluster sizes, the elements of **D** are sampled less efficiently than the other parameters.

Interpretation of the estimates beyond the broad direction of impact, however, is complicated by the nonlinearity in the model and the interactions between the covariates. For example, the income and child variables enter the model in such a way as to make it difficult to disentangle and evaluate their effects. For this reason, Figure 8 presents the average effects for certain changes in the child and income covariates. More specifically, the figure presents the average effects of three hypothetical scenarios: first, the effect of an additional birth in period 1 (i.e., having an additional child age 0–2 in periods 1–3, who grows to become a child age 3–5 in periods 4 and 5), second, the effect of an additional child age 3–5 in periods 1–3, and third, doubling of the husband's earnings. Figure 8 shows that there is a negative overall effect of preschool children on labor supply, which is noticeably stronger for children age 0–2 than

Table 7. Parameter Estimates for Model $\mathcal{M}_3$, Along With 95% Confidence Intervals and Inefficiency Factors From 15,000 MCMC Iterations

Parameter	*Covariate*	*Mean*	*SD*	*Median*	*Lower*	*Upper*	*Ineff*
δ	RACE	.170	.080	.169	.014	.329	7.012
	EDU	.087	.015	.086	.057	.117	23.189
	ln(INC)	−.190	.048	−.189	−.286	−.098	16.484
γ	$\bar{y}_{i0}$	1.371	.173	1.365	1.047	1.724	27.802
	CH2	.142	.312	.144	−.479	.747	5.414
	(CH2) ($\bar{y}_{i0}$)	−.245	.161	−.248	−.556	.077	19.356
	(CH2) ($\overline{\ln(\text{INC}_i)}$)	−.135	.093	−.135	−.318	.046	6.230
	CH5	.868	.273	.867	.339	1.416	8.358
	(CH5) ($\bar{y}_{i0}$)	−.351	.127	−.350	−.606	−.103	14.530
	(CH5) ($\overline{\ln(\text{INC}_i)}$)	−.221	.081	−.221	−.380	−.063	8.139
ϕ	$y_{i,t-1}$	1.213	.071	1.213	1.072	1.348	15.863
	$y_{i,t-2}$	.445	.071	.445	.308	.581	11.470
vech(**D**)		.540	.129	.528	.319	.828	38.481
		−.043	.096	−.043	−.243	.133	45.999
		.137	.071	.119	.046	.319	45.617
		−.151	.085	−.138	−.347	−.019	43.454
		.017	.049	.011	−.066	.136	45.551
		.158	.086	.135	.047	.366	46.355
τ^2		.017	.006	.016	.009	.030	5.473

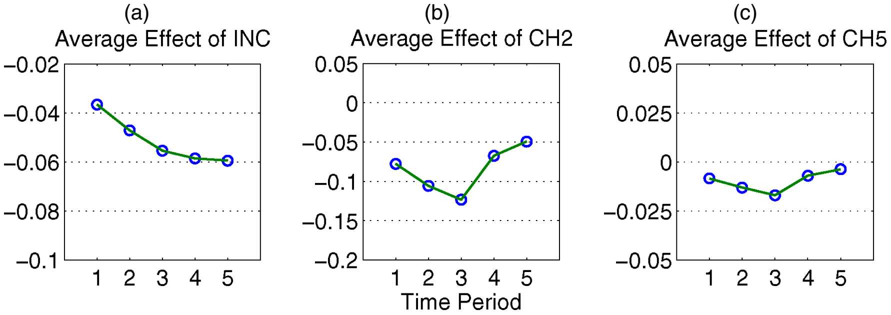

Figure 8. The Average Effect of (a) Doubling a Husband's Permanent Income, (b) Having an Additional Child Age 0–2 in Periods 1–3 and Age 3–5 in Periods 4 and 5, and (c) an Additional Child Age 3–5 in Periods 1–3 (c).

for children age 3–5 (cf. Hyslop 1999). The figure also shows that although husband's earnings affect a woman's probability of employment in a theoretically predicable direction, increases in earnings must be quite large to induce any economically significant reduction in participation. In some situations, we are interested in the effect of a given policy intervention on different segments of the population. To explore this issue, we can compute and compare the average covariate effects for various subsamples, as discussed in Section 4. Figure 9 shows the average covariate effects for two subsamples based on their initial conditions: women who have worked in both presample periods and women who have worked in the first, but not the second presample period (the other two categories are not shown). The figure reveals considerable differences in the impact of the covariates on labor force choices in these two groups and raises some interesting questions for future research.

8. CONCLUDING REMARKS

This article has considered the Bayesian analysis of hierarchical semiparametric models for binary panel data with state dependence, serially correlated errors, and multidimensional heterogeneity correlated with the covariates and initial conditions. New, computationally efficient MCMC algorithms have been developed for simulating the posterior distribution, estimating the marginal likelihood, and evaluating the average covariate effects. The techniques rely on the framework of Albert and Chib (1993) and a proper Markov process smoothness prior on the unknown function. A simulation study has shown that the methods performs well. An application involving a dynamic semiparametric model of women's labor force participation illustrated that the model and the estimation methods are practical and can uncover interesting features in the data. Formal Bayesian model choice methods allowed us to distinguish among the effects of state dependence, serial correlation, and heterogeneity, and to compare parametric and semiparametric models.

One benefit of the model considered herein is that it can be inserted as a component in a larger hierarchical model (e.g., a treatment model or a model with incidental truncation). The general method is also applicable to panels of continuous and censored data. We intend to explore the effectiveness of such approaches in future work.

[Received November 2004. Revised June 2005.]

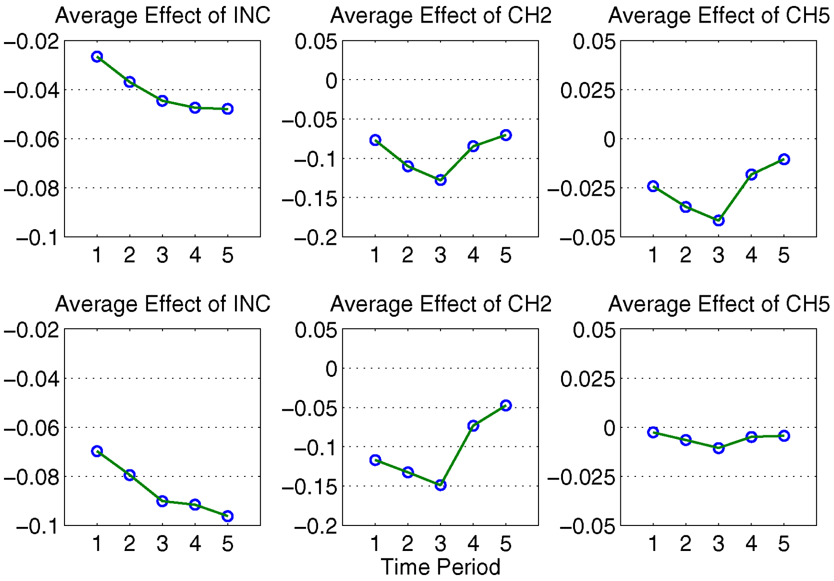

Figure 9. Average Covariate Effects for Two Subsamples: Women Who Worked in Both Presample Periods (first row) and Women Who Worked in the First, but Not the Second, Presample Period (second row).

REFERENCES

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

Altman, N. S. (1990), "Kernel Smoothing of Data With Correlated Errors," *Journal of the American Statistical Association*, 85, 749–759.

Avery, R., Hansen, L., and Hotz, V. (1983), "Multiperiod Probit Models and Orthogonality Condition Estimation," *International Economic Review*, 24, 21–35.

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

Börsch-Supan, A., and Hajivassiliou, V. (1993), "Smooth Unbiased Multivariate Probability Simulators for Maximum Likelihood Estimation of Limited Dependent Variable Models," *Journal of Econometrics*, 58, 347–368.

Chib, S. (1993), "Bayes Estimation of Regressions With Autoregressive Errors: A Gibbs Sampling Approach," *Journal of Econometrics*, 58, 275–294.

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

Chib, S., and Carlin, B. (1999), "On MCMC Sampling in Hierarchical Longitudinal Models," *Statistics and Computing*, 9, 17–26.

Chib, S., and Greenberg, E. (1994), "Bayes Inference in Regression Models With ARMA (p, q) Errors," *Journal of Econometrics*, 64, 183–206.

——— (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.

Diggle, P., and Hutchinson, M. (1989), "On Spline Smoothing With Autocorrelated Errors," *Australian Journal of Statistics*, 31, 166–182.

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.

Gelman, A., Carlin, B., Stern, H., and Rubin, D. (1995), *Bayesian Data Analysis*, New York: Chapman & Hall.

Geweke, J. (1991), "Efficient Simulation From the Multivariate Normal and Student-*t* Distributions Subject to Linear Constraints," in *Computing Science and Statistics: Proceedings of the Twenty-Third Symposium on the Interface*, ed. E. M. Keramidas, Fairfax, VA: Interface Foundation of North America, pp. 571–578.

Hamilton, J. D. (1994), *Time Series Analysis*, Princeton, NJ: Princeton University Press.

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

Harvey, A. C. (1981), *The Econometric Analysis of Time Series*, Oxford, U.K.: Phillip Allen.

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

Hastings, W. K. (1970), "Monte Carlo Sampling Methods Using Markov Chains and Their Applications," *Biometrika*, 57, 97–109.

Heckman, J. (1981), "Heterogeneity and State Dependence," in *Studies in Labor Markets*, ed. S. Rosen, Chicago: University of Chicago Press, pp. 91–131.

Hyslop, D. (1999), "State Dependence, Serial Correlation and Heterogeneity in Intertemporal Labor Force Participation of Married Women," *Econometrica*, 67, 1255–1294.

Kass, R., and Raftery, A. (1995), "Bayes Factors," *Journal of the American Statistical Association*, 90, 773–795.

Karcher, P., and Wang, Y. (2001), "Generalized Nonparametric Mixed Effects Models," *Journal of Computational and Graphical Statistics*, 10, 641–655.

Keane, M. (1994), "A Computationally Practical Simulation Estimator for Panel Data," *Econometrica*, 62, 95–116.

Kohn, R., Ansley, C. F., and Tharm, D. (1991), "The Performance of Cross-Validation and Maximum Likelihood Estimators of Spline Smoothing Parameters," *Journal of the American Statistical Association*, 66, 1042–1050.

Koop, G., and Poirier, D. J. (2004), "Bayesian Variants of Some Classical Semiparametric Regression Techniques," *Journal of Econometrics*, 123, 259–282.

Lin, X., and Carroll, R. (2001), "Semiparametric Regression for Clustered Data Using Generalized Estimating Equations," *Journal of the American Statistical Association*, 96, 1045–1056.

Lin, X., and Zhang, D. (1999), "Inference in Generalized Additive Mixed Models Using Smoothing Splines," *Journal of the Royal Statistical Society*, Ser. B, 61, 381–400.

Mundlak, Y. (1978), "On the Pooling of Time Series and Cross-Section Data," *Econometrica*, 46, 69–85.

O'Hagan, A. (1994), *Kendall's Advanced Theory of Statistics: Bayesian Inference*, Vol. 2B, London: Edward Arnold.

Opsomer, J. D., Wang, Y., and Yang, Y. (2001), "Nonparametric Regression With Correlated Errors," *Statistical Science*, 16, 134–153.

Ortega, J. (1987), *Matrix Theory*, New York: Plenum Press.

Shiller, R. (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., Wong, C.-M., and Kohn, R. (1998), "Additive Nonparametric Regression With Autocorrelated Errors," *Journal of the Royal Statistical Society*, Ser. B, 311–331.

Tierney, L. (1994), "Markov Chains for Exploring Posterior Distributions" (with discussion), *The Annals of Statistics*, 22, 1701–1762.

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.

Wang, Y. (1998), "Smoothing Spline Models With Correlated Random Errors," *Journal of the American Statistical Association*, 93, 341–348.

Wood, S., and Kohn, R. (1998), "A Bayesian Approach to Robust Binary Nonparametric 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.

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