Semiparametric Hierarchical Bayes Analysis of Discrete Panel Data with State Dependence and Serial Correlation

Siddhartha Chib*
Washington University in St. Louis

Ivan Jeliazkov†
University of California-Irvine

This version: April, 2004

Abstract

In this paper we consider the analysis of semiparametric models for binary panel data with state dependence and serial correlation. A hierarchical approach is used in addressing heterogeneity, dealing with the initial conditions, and incorporating correlation between the covariates and the random effects. We consider a semiparametric specification in which a Markov process smoothness prior is used to model an unknown regression function. The paper presents new computationally efficient Markov chain Monte Carlo estimation algorithms. Simulation results suggest that the methods perform well. In addition to estimation, we address the problem of model choice and compare competing parametric and semiparametric specifications. Moreover, we present a framework for calculating the average covariate effects, which deals with the nonlinearity and dynamic structure of the model. The techniques of this paper are used to study the intertemporal labor force participation decisions of a panel of 1545 married women. In this application, the data support a semiparametric model with multiple sources of heterogeneity and multi-lag state dependence.

Keywords: Average covariate effects; Bayes factor; Bayesian model comparison; Correlated binary data; Clustered data; Labor force participation; Marginal likelihood; Markov chain Monte Carlo (MCMC); Markov process priors; Nonparametric estimation; Partially linear model.
JEL Classification: C11, C14, C15, C35, C51, C52, C63, J22.

1 Introduction

This article discusses techniques for analyzing semiparametric models for dynamic binary panel data, and applies them to study women's labor force participation choices using data from the Panel Study of Income Dynamics (PSID). We adopt a hierarchical Bayesian approach to integrate and extend a number of modeling and estimation techniques and provide a flexible specification and inferential framework for models with multidimensional heterogeneity, general dynamic dependence, and nonparametric functional form. In addition, we address the problem

*Address for correspondence: John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Drive, St. Louis, MO 63130. E-mail: chib@wustl.edu.

†Address for correspondence: Department of Economics, University of California-Irvine, 3151 Social Science Plaza, Irvine, CA 92697-5100. E-mail: ivan@uci.edu.

of model choice and propose a simulation-based approach for evaluation of the average covariate effects. The former enables the formal comparison of semiparametric to parametric models; the latter provides interpretability of the estimates.

Let y_{it} be the binary response of interest, where the indices i and t ($i = 1, ..., n$, $t = 1, ..., T_i$) refer to units (individuals, firms, countries, etc.) and time, respectively. We consider a dynamic partially linear binary choice model where y_{it} depends parametrically on the covariate vectors $\mathbf{x}_{it}^*$ and $\mathbf{w}_{it}$, and nonparametrically on the covariate s_{it} in the form

$$y_{it} = 1\{\mathbf{x}_{it}^{*\prime}\delta + \mathbf{w}_{it}^{\prime}\beta_i + g\left(s_{it}\right) + \phi_1 y_{i,t-1} + ... + \phi_J y_{i,t-J} + \varepsilon_{it} > 0\}, \tag{1}$$

where $1\{\cdot\}$ is an indicator function, δ and β_i are vectors of common and unit-specific (random) effects, respectively, $\phi_1, ..., \phi_J$ are lag coefficients, $g\left(\cdot\right)$ is an unknown function, and ε_{it} is a possibly serially correlated error term. Inclusion of a general number of lags $y_{i,t-1}, ..., y_{i,t-J}$ in (1) captures the notion of "state dependence", where the probability of response may depend on past occurrences because of altered preferences, trade-offs, or constraints.[1] The individual effects β_i account for heterogeneity in the effects of the covariates $\mathbf{w}_{it}$, but modeling heterogeneity is also essential in guarding against the emergence of "spurious state dependence", where temporal pseudodependence occurs because history may serve as a proxy for the unobserved unit-specific propensities to experience the event (Heckman 1981, Hsiao 1986). As discussed in Section 2.1, we allow the heterogeneity to depend on the initial observations and the covariates.

The semiparametric model specified in (1) is new in the analysis of binary panel data. Much of the recent work on dynamic binary panel data models (Hyslop (1999), Honoré and Kyriazidou (2000), Klaassen and Magnus (2001)), has emphasized parametric specifications. An exception is the semiparametric extension discussed in Honoré and Kyriazidou (2000), although they do not consider serial correlation in the errors and restrict heterogeneity to the intercept. In contrast, the approach taken here can simultaneously accommodate a partially linear index function, general state dependence, serial correlation, and multiple correlated random effects, but in the setting of known disturbance and random effect distributions. The overall approach here can be viewed as an extension of the Bayesian nonparametric modeling of Wahba (1978), Shiller

[1] Typical examples include applied problems such as the analysis of investment, transportation, health status and illness recurrence, fertility decisions, accident occurrence, employment, and labor union participation.

(1984), Wood and Kohn (1998), and Fahrmeir and Lang (2001) to a dynamic panel data model. For an overview of partially linear models for cross-section data, see DiNardo and Tobias (2001) and Yatchew (1998).

The presence of the nonparametric function in the binary response model in (1) raises a number of challenges for estimation due to the intractability of the likelihood function. Many of the problems have been largely overcome in the Bayesian context by Wood and Kohn (1998), Shively *et al.* (1999), and Wood *et al.* (2002), based on the Markov chain Monte Carlo (MCMC) framework for dealing with binary data proposed in Albert and Chib (1993). One open problem, however, is the question of model comparison in the semiparametric setting. Previous work in settings with cross-sectional binary data has relied on measures such as the AIC and the BIC, but their computation is infeasible in our context. We extend that literature by showing how the problem can be tackled formally through the calculation of marginal likelihoods and Bayes factors. Another open problem is the analysis of semiparametric models with serially correlated errors. The issue has been studied in Diggle and Hutchinson (1989), Altman (1990), and Smith *et al.* (1998), however their estimation algorithms are $\mathcal{O}(N^3)$, where $N = \sum_{i=1}^{n} T_i$ is the total number of observations in the sample. Here we propose a new $\mathcal{O}(N)$ algorithm for estimation of the nonparametric function when the errors are correlated. The computations are carried out by a simulation algorithm that exploits the panel structure of the data to orthogonalize the errors before simulating the unknown functions. Yet another open problem is the interpretation of the coefficients in the setting of such non-linear models. We develop a predictive framework for calculating and describing the average effect of a given covariate x on the probability of response, both contemporaneously and over time.

We apply our model and methods to the setting of Hyslop (1999), who studied women's labor force participation using dynamic probit and linear probability models. While women's labor supply in the U.S. has increased on the intensive (hours of work) margin, the most important development has been the increase on the extensive (participation) margin. In the case of married women, Goldin (1989) reports that the seven-fold increase in participation since the 1920's has not been associated with a substantial increase in average work experience among employed married women. In a recent survey of the labor supply literature, Heckman (1993)

underscores the importance of participation decisions and concludes that much of the elasticity of estimated labor supply functions comes through entry and exit decisions. One of the salient features of these participation choices, which is also essential for policy considerations, is their state dependence. However, Hyslop (1999) notes that despite its importance, the intertemporal labor supply behavior of women remains one of the most difficult areas of applied labor supply research. We augment Hyslop's findings by estimating several parametric and semiparametric specifications that differ in their dynamics, heterogeneity, and functional form. These models are then compared to gauge the appropriateness of the underlying statistical specifications. The results from our regressions strengthen Hyslop's (1999) finding that participation is characterized by significant state dependence; properly accounting for those dynamics is crucial in eliciting the effects of the other covariates (e.g. fertility decisions, race, husband's income) on participation. We also find support for the presence of heterogeneity in the effect of children on the mother's labor supply (Angrist and Evans 1998). This heterogeneity is correlated with the husband's income. The analysis also indicates that age is non-linearly related to labor force participation.

The rest of the paper is organized as follows. In Section 2 we complete the statistical model. In Section 3 we present our MCMC based fitting methods and in Section 4, we show how the average effects of the covariates on the probability of response are calculated. Section 5 is concerned with the comparison of the model to various alternatives, based on marginal likelihoods and Bayes factors. Section 6 presents a detailed simulation study of the performance of the estimation method. In Section 7 we study the intertemporal labor force participation of a panel of married women. Concluding remarks are presented in Section 8.

2 Statistical Modeling

2.1 Hierarchical Modeling of the Unobserved Effects

To explain the heterogeneity modeling, 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

$$z_{it} = \mathbf{x}_{it}^{*\prime}\delta + \mathbf{w}_{it}^{\prime}\beta_i + g(s_{it}) + \phi_1 1\{z_{i,t-1} > 0\} + ... + \phi_J 1\{z_{i,t-J} > 0\} + \varepsilon_{it}$$

where $y_{it} = 1\{z_{it} > 0\}$ and $\varepsilon_{it} \sim \mathcal{N}(0,1)$ for all i and t.[2] Let $\mathbf{y}_{i0} \equiv (y_{i,-J+1}, ..., y_{i0})'$ be the J-vector of initial observations for subject i, $\mathbf{y}_i \equiv (y_{i1}, ..., y_{iT_i})$ denote the remaining T_i observations in the ith cluster, and define the lag vectors

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

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

$$\mathbf{z}_i = \mathbf{X}_i^* \delta + \mathbf{W}_i \beta_i + \mathbf{g}_i + \mathbf{L}_i \phi + \varepsilon_i \tag{2}$$

where $\mathbf{z}_i = (z_{i1}, ..., z_{iT_i})'$, $\mathbf{X}_i^* = \left(\mathbf{x}_{i1}^*, ..., \mathbf{x}_{iT_i}^*\right)'$, $\mathbf{W}_i = (\mathbf{w}_{i1}, ..., \mathbf{w}_{iT_i})'$, $\mathbf{g}_i = (g(s_{i1}), ..., g(s_{iT_i}))'$, $\mathbf{s}_i = (s_{i1}, ..., s_{iT_i})'$, $\mathbf{L}_i = (\mathbf{y}_{i,-1}, ..., \mathbf{y}_{i,-J})$, $\phi = (\phi_1, ..., \phi_J)'$, and $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT_i})'$ is distributed as multivariate normal with mean zero and an identity covariance matrix.[3] Subsequently, we will also consider a model in which the errors are serially correlated.

Now consider the modeling of the unobserved effects. Based on Mundlak (1978), Chamberlain (1984), and Wooldridge (2000), we assume that the distribution of the q-vector β_i is Gaussian with mean value that depends on the initial observations $\mathbf{y}_{i0}$ and the covariates. In particular, we let

$$\beta_i | \mathbf{y}_{i0}, \mathbf{X}_i^*, \mathbf{W}_i, \mathbf{s}_i, \gamma, \mathbf{D} \sim \mathcal{N}(\mathbf{A}_i \gamma, \mathbf{D}), \quad i = 1, ..., n, \tag{3}$$

or equivalently

$$\beta_i = \mathbf{A}_i \gamma + \mathbf{b}_i, \quad \mathbf{b}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{D}), \quad i = 1, ..., n, \tag{4}$$

where the matrix $\mathbf{A}_i$ can be defined quite flexibly, given the specifics of the problem at hand. In the simplest case where $\mathbf{W}_i$ does not include an intercept and β_i is independent of the covariates, a parsimonious way of modeling the dependence of β_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 $\bar{y}_{i0} = (1/J) \sum_{j=1-J}^{0} y_{ij}$ is the mean of the entries in $\mathbf{y}_{i0}$. Moreover, the matrix $\mathbf{A}_i$ may also contain within-cluster means of a subset of covariates – those which are suspected of being correlated with the random effects for each cluster (Mundlak 1978). If $\bar{\mathbf{r}}_{ij}$

[2] In the pre-sample ($t = -J+1, ..., 0$), the latent data $\{z_{it}\}$ are not modeled and, for the purposes of this discussion, can simply be taken to equal the corresponding pre-sample $\{y_{it}\}$.

[3] In the above, the vectors $\mathbf{x}_{it}^*$ and $\mathbf{w}_{it}$ contain two disjoint sets of covariates and because $g(\cdot)$ is unrestricted, $\mathbf{x}_{it}^*$ does not include an intercept or the covariate s_{it}, although those may be included in $\mathbf{w}_{it}$.

$(j = 1, ..., q)$ denotes the vector of such covariate means, the general $\mathbf{A}_i$ may be written as

$$\mathbf{A}_i = \begin{pmatrix} 1 & \bar{y}_{i0} & \bar{\mathbf{r}}'_{i1} & & & \\ & & & \ddots & & \\ & & & & 1 & \bar{y}_{i0} & \bar{\mathbf{r}}'_{iq} \end{pmatrix} \tag{5}$$

where the dimension of $\bar{\mathbf{r}}_{ij}$ depends on how many column averages of $\mathbf{X}_i^*$, $\mathbf{W}_i$, and $\mathbf{s}_i$ were allowed to influence the respective random effect. The mean $E(\beta_i|\mathbf{y}_{i0}, \mathbf{X}_i^*, \mathbf{W}_i, \mathbf{s}_i, \gamma, \mathbf{D})$ need not necessarily be modeled as a linear function. In such cases, modeling can rely on higher order terms or other summaries of the covariates to the matrix $\mathbf{A}_i$.

It can be seen that upon utilizing (4), equation (2) can be equivalently expressed as

$$\mathbf{z}_i = \mathbf{X}_i\beta + \mathbf{g}_i + \mathbf{W}_i\mathbf{b}_i + \varepsilon_i \tag{6}$$

where

$$\mathbf{X}_i = \begin{pmatrix} \mathbf{X}_i^* & \mathbf{W}_i\mathbf{A}_i & \mathbf{L}_i \end{pmatrix}$$

and

$$\beta = \begin{pmatrix} \delta' & \gamma' & \phi' \end{pmatrix}'.$$

We note that neither $\mathbf{s}_i$ nor an intercept should be present in the matrix $\mathbf{X}_i$, which achieves the dual purpose of allowing for useful modeling of the inter-cluster heterogeneity, but at the same time resolves the identification problem under a general, unrestricted $g(\cdot)$. In more general specifications where a random intercept is included in the model, one has to adjust $\mathbf{A}_i$ for identification purposes. If the random intercept is the ith column of $\mathbf{W}_i$, the column of $\mathbf{A}_i$ which 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 $\mathbf{W}_i$, the column of $\mathbf{A}_i$ which 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 $\mathbf{X}_i^*$ 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 (since $\mathbf{W}_i\mathbf{A}_i$ 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 (semi-conjugate) prior densities for the model parameters β and $\mathbf{D}$. Gaussian priors are used to summarize the

prior information about the k-vector β, while a Wishart prior is used for the $q \times q$ matrix $\mathbf{D}^{-1}$:

$$\beta \sim \mathcal{N}(\beta_0, \mathbf{B}_0), \quad \mathbf{D}^{-1} \sim \mathcal{W}(r_0, \mathbf{R}_0). \tag{7}$$

The specification of the prior for the function $g(\cdot)$ is discussed next.

2.2 The Prior on $g(\cdot)$

We place a Markov process smoothness prior on the function $g(\cdot)$. A range of similar smoothness priors can be found in the literature on nonparametric modeling; for some specific examples, see Shiller (1973, 1984), Wahba (1978), Silverman (1985), Besag *et al.* (1995), Fahrmeir and Tutz (1997, Chapter 8), Fahrmeir and Lang (2001), and Koop and Poirier (2004). While these priors differ in the specifics, they all rest on the idea that local variation in the function, as measured by changes in its derivatives or divided differences, should not be too large. The roots for this method can be traced back to Whittaker's (1923) penalized least squares criterion, where the aim is to strike a balance between a good fit and a smooth regression function. Some discussion on the similarities between these priors is offered in Shiller (1984), Silverman (1985), Besag *et al.* (1995), Fahrmeir and Tutz (1997), and Fahrmeir and Lang (2001). It is also interesting to note that the Bayesian and non-Bayesian approaches to nonparametric modeling are quite similar because there is a direct correspondence between Bayesian priors on the unknown function and the penalties that frequentists specify in penalized likelihood estimation.

In our context, suppose that the N observations in the covariate vector $\mathbf{s}$ determine the $m \times 1$ *design point vector* $\mathbf{v}$ with entries equal to the *unique ordered* values of $\mathbf{s}$ with

$$v_1 < ... < v_m,$$

and with

$$\mathbf{g} = (g(v_1), ..., g(v_m))' = (g_1, ..., g_m)', \tag{8}$$

being 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 $\mathbf{g}$. In our implementation, the function evaluations are modeled as resulting from the realization of a second order Markov process, with the specification aimed at penalizing rough functions $g(\cdot)$.

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

$$g_t = \left(1 + \frac{h_t}{h_{t-1}}\right) g_{t-1} - \frac{h_t}{h_{t-1}} g_{t-2} + u_t,\ u_t \sim N\left(0, \tau^2 h_t\right), \tag{9}$$

where τ^2 is a smoothness parameter. Small values of τ^2 produce smoother functions; 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 (e.g. see 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.

To complete the specification of the smoothness prior, we provide a distribution for the initial states of the random walk process

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

where $\mathbf{G}_0$ is a 2×2 symmetric positive definite matrix. The prior on the initial conditions (10) induces a prior on linear functions of $\mathbf{v}$, which 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 (9) in expectation (to eliminate u_t which is the source of the nonlinearity), starting with initial states as specified in (10). Thus, conditional on g_1 and g_2, the mean of the Markov process in (9) 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 that is directly related to the distribution in (10) in a one-to-one mapping. This is useful in setting the prior parameters g_{10}, g_{20}, and $\mathbf{G}_0$. For example, if $v_1 = 0$, $v_2 = 1$, and $\tau^2 = 1$, the unconditional mean of the Markov process under $g_{10} = 0$, $g_{20} = 0$, and $\mathbf{G}_0 = \mathbf{I}$, is equivalent to a prior of $\mathcal{N}(0, 1)$ on the intercept and $\mathcal{N}(0, 2)$ on the slope parameter in a corresponding linear model.

The directed Markovian structure of the random walk prior specified by (9) and (10) emphasizes a local smoothness penalty. An equivalent (global) smoothness prior for $\mathbf{g}$ results after

rewriting the Markov process in a random field form. To see this, note that after defining

$$\mathbf{H} = \begin{pmatrix} 1 & & & & \\ & 1 & & & \\ \frac{h_t}{h_{t-1}} & -\left(1+\frac{h_t}{h_{t-1}}\right) & 1 & & \\ & \ddots & \ddots & \ddots & \\ & & \frac{h_m}{h_{m-1}} & -\left(1+\frac{h_m}{h_{m-1}}\right) & 1 \end{pmatrix},$$

and

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

the global smoothness representation of the second order Markov process prior equivalent to (9) and (10) becomes

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

where $\mathbf{g}_0 = \mathbf{H}^{-1}\tilde{\mathbf{g}}$, with $\tilde{\mathbf{g}} = (g_{10}, g_{20}, 0, ..., 0)'$, and the *penalty matrix* $\mathbf{K}$ is given by $\mathbf{K} = \mathbf{H}'\mathbf{\Sigma}^{-1}\mathbf{H}$. Equivalently, $\mathbf{g}_0$ can be derived by taking recursive expectations of (9) starting with the mean in (10), and as argued above, the points in $\mathbf{g}_0$ will form a straight line.

A key feature of the prior in (11) is that it is proper. This simple innovation offers an important refinement on much of the literature on smoothness priors for nonparametric function estimation where, in contrast, partially improper priors and reduced rank penalty matrices $\mathbf{K}$ are used. Since improper priors preclude the possibility for formal finite sample model comparison using marginal likelihoods and Bayes factors, our approach removes an important impediment to formal Bayesian model selection. We discuss the approach to model selection in Section 4 below. Since the prior on $\mathbf{g}$ is defined conditional of the hyperparameter τ^2, in the next level of the hierarchy we specify the prior distribution

$$\tau^2 \sim \mathcal{IG}\left(\frac{\nu_0}{2}, \frac{\delta_0}{2}\right). \tag{12}$$

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 *et al.* (1995, Appendix A)). How the choice of these parameters will affect the estimated $\mathbf{g}$ will depend on the other sources of variance in the model. Some intuition about this can be gained by considering the sampling algorithm 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 $\mathbf{g}_0$. Thus, τ^2 competes with the inverse of the error variance, which weighs the function $\mathbf{g}$ that maximizes the fit in the likelihood (thus illustrating the trade-off between smoothness and good fit that was discussed in Whittaker (1923)).

We conclude the discussion on Markov process priors by making two remarks. First, from an estimation point of view, it is important to note that the penalty matrix $\mathbf{K}$ is banded. This fact is of considerable practical utility, as manipulations involving banded matrices take $\mathcal{O}(m)$ operations, rather than the usual $\mathcal{O}(m^3)$ for inversions or $\mathcal{O}(m^2)$ for multiplication by a vector. Given that m may be large (potentially as large as the total number of observations N in the panel) this has important ramifications for the numerical efficiency of the estimation procedure. Second, Markov process priors are conceptually simple and easily adaptable to different orders, enabling them to match problem-specific tasks more closely (Besag *et al.* (1995), Fahrmeir and Lang (2001)). For example, a simple first order Markov process prior $g_t = g_{t-1} + u_t$ penalizes abrupt jumps $g_t - g_{t-1}$ between successive states of the random walk process, while higher order priors embody more complex notions of "smoothness" related to the rates of change in the function. Such priors share many similar features and are easily specified using the general ideas outlined above.

3 Estimation

We first address the Bayesian estimation of the semiparametric model with unobserved heterogeneity and state dependence. The estimation algorithm is based on MCMC simulation of the posterior distribution (see Chib (2001) for details of these methods). We then provide a new method for models with serial correlation in the errors.

3.1 Model with State Dependence

From (6) we see that after marginalizing β_i using the distribution in (3) that the latent $\mathbf{z}_i$ can be expressed as

$$\mathbf{z}_i = \mathbf{X}_i\beta + \mathbf{g}_i + \mathbf{u}_i \tag{13}$$

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

$$\Pr(\mathbf{y}_i | \mathbf{y}_{i0}, \beta, \mathbf{g}_i, \mathbf{D}) = \int_{B_{iT_i}} \cdots \int_{B_{i1}} \mathcal{N}(\mathbf{z}_i | \mathbf{X}_i \beta + \mathbf{g}_i, \mathbf{V}_i) d\mathbf{z}_i \,, \tag{14}$$

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.

However, estimation of the model in the Bayesian context is possible under the framework of Albert and Chib (1993) by including the latent $\{\mathbf{z}_i\}$ as part of the unknowns of the model. To describe this approach we stack the observations in (6) across all subjects in the sample as

$$\mathbf{z} = \mathbf{X}\beta + \mathbf{Q}\mathbf{g} + \mathbf{W}\mathbf{b} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{I}) \,, \tag{15}$$

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

$$\mathbf{X} = \begin{bmatrix} \mathbf{X}_1 \\ \vdots \\ \mathbf{X}_n \end{bmatrix},$$

the block-diagonal matrix

$$\mathbf{W} = \begin{bmatrix} \mathbf{W}_1 & & \\ & \ddots & \\ & & \mathbf{W}_n \end{bmatrix},$$

and where $\mathbf{Q}$ 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 $\mathbf{v}$, and all remaining elements are zeros, so that $\mathbf{s} = \mathbf{Q}\mathbf{v}$.

The MCMC algorithm described below is derived form (13) and (15), and is based on an efficient blocking scheme proposed by Chib and Carlin (1999). They note that because β and $\{\mathbf{b}_i\}$ are correlated by construction, sampling them in two separate blocks from their full conditional densities results in a slowly mixing and inefficient algorithm. In contrast, there is a significant improvement in the efficiency of the Markov chain sampler when the fixed and random effects are sampled in one block. This is done by using (13) to sample β from a conditional density that does not depend on $\{\mathbf{b}_i\}$, followed by drawing $\{\mathbf{b}_i\}$ from its full conditional density. We note that with data augmentation, there are now two sets of latent variables in the sampler, $\{\mathbf{z}_i\}$ and

$\{\mathbf{b}_i\}$, and that the algorithm below subsumes an algorithm for the estimation of a simpler, fully parametric version of the model.

Algorithm 1 ***Gaussian State Dependence Model: MCMC Implementation***

1. Sample $\{\mathbf{z}_i\}|\mathbf{y}, \mathbf{D}, \mathbf{g}, \beta$ marginal of $\{\mathbf{b}_i\}$ by drawing for $i \leq n$, $t \leq T_i$

$$z_{it} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 1 \\ \mathcal{TN}_{(-\infty,0]}(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 0 \end{cases}$$

   where $\mathcal{TN}_{(a,b)}(\mu_{it}, v_{it})$ is a normal distribution truncated to the interval (a, b) with mean $\mu_{it} = E(z_{it}|\mathbf{z}_{i\setminus t}, \beta, \mathbf{g}_i, \mathbf{V}_i)$ and variance $v_{it} = Var(z_{it}|\mathbf{z}_{i\setminus t}, \beta, \mathbf{g}_i, \mathbf{V}_i)$, and where $\mathbf{V}_i = \mathbf{I} + \mathbf{W}_i\mathbf{D}\mathbf{W}_i'$.

2. Sample $\beta, \{\mathbf{b}_i\}|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g}_i$ in one block by drawing

   (a) $\beta|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g} \sim \mathcal{N}(\hat{\beta}, \mathbf{B})$, where $\hat{\beta} = \mathbf{B}(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^n \mathbf{X}_i'\mathbf{V}_i^{-1}(\mathbf{z}_i - \mathbf{g}_i))$ and $\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{X}_i'\mathbf{V}_i^{-1}\mathbf{X}_i)^{-1}$ are the usual updates based on the complete data;

   (b) $\mathbf{b}_i|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \beta, \mathbf{g} \sim \mathcal{N}(\hat{\mathbf{b}}_i, \mathbf{B}_i)$ with $\hat{\mathbf{b}}_i = \mathbf{B}_i\mathbf{W}_i'(\mathbf{z}_i - \mathbf{X}_i\beta - \mathbf{g}_i)$ and $\mathbf{B}_i = (\mathbf{D}^{-1} + \mathbf{W}_i'\mathbf{W}_i)^{-1}$ for $i = 1, ..., n$.

3. Sample $\mathbf{D}^{-1}|\{\mathbf{b}_i\} \sim \mathcal{W}\left\{r_0 + n, \left(\mathbf{R}_0^{-1} + \sum_{i=1}^n \mathbf{b}_i\mathbf{b}_i'\right)^{-1}\right\}$.

4. Sample $\mathbf{g}|\mathbf{y}, \beta, \{\mathbf{b}_i\}, \tau^2, \{z_{it}\} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$, where $\mathbf{G} = \left(\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{Q}\right)^{-1}$ and where $\hat{\mathbf{g}} = \mathbf{G}\left(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'\left(\mathbf{z} - \mathbf{X}\beta - \mathbf{W}\mathbf{b}\right)\right)$. Remark 1 below presents an important note on the sampling in this step.

5. Sample $\tau^2|\mathbf{g} \sim \mathcal{IG}\left(\frac{\nu_0+m}{2}, \frac{\delta_0+(\mathbf{g}-\mathbf{g}_0)'\mathbf{K}(\mathbf{g}-\mathbf{g}_0)}{2}\right)$.

Remark 1. In sampling $\mathbf{g}$, one should note that $\mathbf{Q}'\mathbf{Q}$ is a diagonal matrix whose jth diagonal entry equals the number of values in $\mathbf{s}$ corresponding to the design point v_j. Since $\mathbf{K}$ 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 $\mathbf{G}^{-1}\hat{\mathbf{g}} = \left(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}'\left(\mathbf{z} - \mathbf{X}\beta - \mathbf{W}\mathbf{b}\right)\right)$, 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 obtain a random draw from $\mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$

efficiently, 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 $\mathbf{x}$, one obtains a draw $\mathbf{g} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$.

We note that the MCMC approach to estimating τ^2 in this hierarchical model overcomes some of the difficulties associated with cross-validation and generalized cross-validation (Craven and Wahba 1979). There are two main advantages of the MCMC approach. First, it can be applied to both continuous and binary data (with the latter being the main focus of this paper), while cross-validation techniques are mainly applicable to continuous data and are infeasible in the current setup. Second, MCMC estimation accounts fully for parameter uncertainty, unlike plug-in approaches, which do not account for the variability due to estimating the smoothing parameters. An important alternative to the above methods is the maximum integrated likelihood approach to determining τ^2 (Kohn *et al.* 1991), but it is also infeasible in our setting because of the intractability of the likelihood function and the high-dimensional integration (over a space of dimension equal to the sample size) that is required by that approach.

Critical to the feasibility of the estimation approach is the existence of efficient MCMC algorithms that can quickly explore the posterior distribution. Algorithm 1 provides one such estimation method for the model with state dependence. However, the presence of serially correlated errors presents other challenges.

3.2 Model with Dependent Errors

Suppose now that the errors follow a zero-mean stationary pth order autoregressive, AR(p), process

$$\varepsilon_{it} = \rho_1 \varepsilon_{t,t-1} + ... + \rho_p \varepsilon_{i,t-p} + v_{it}\,, \tag{16}$$

where v_{it} is independent $\mathcal{N}(0,1)$. The process in (16) can equivalently be expressed in terms of a polynomial in the backshift operator L as $\rho(L)\,\varepsilon_{it} = v_{it}$, where $\rho(L) = 1 - \rho_1 L - ... - \rho_p L^p$ and stationarity is maintained by requiring that all roots of $\rho(L)$ lie outside the unit circle. We clarify that, in contrast, stationarity was not an issue for the state dependence coefficients ϕ because those multiply the binary lags and thus simply serve as intercept shifts.

Previous studies by Diggle and Hutchinson (1989), Altman (1990), and Smith *et al.* (1998), have shown that when the errors are treated as independent when in fact 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 as both are stochastic processes in time. If the serial correlation in the errors is ignored, 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) point out that even if the independent variable is not time, modeling the autocorrelation in the errors gives more efficient nonparametric estimates, as it reduces the effective error variance similarly to the case of parametric regression.

In our longitudinal data setup, for $i = 1, ..., n$ and $t = 1, ..., T_i$, the latent data representation of the model with AR(p) serial correlation is

$$\mathbf{z}_i = \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \mathbf{g}_i + \varepsilon_i,$$

where $y_{it} = 1\{z_{it} > 0\}$, the errors $\varepsilon_i = (\varepsilon_{i1}, ..., \varepsilon_{iT_i})'$ follow the distribution $\varepsilon_i \sim \mathcal{N}(\mathbf{0}, \mathbf{\Omega}_i)$, and $\mathbf{\Omega}_i$ is the $T_i \times T_i$ Toeplitz matrix implied by the autoregressive process. For the general AR(p) case, the matrix $\mathbf{\Omega}_i$ 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 easily be shown (cf Hamilton 1994, Section 3.4) that the autocovariances follow the same pth-order difference equation as does the process itself, i.e. $\varphi_j = \rho_1\varphi_{j-1} + ... + \rho_p\varphi_{j-p}$. The first p values $(\varphi_0, \varphi_1, ..., \varphi_{p-1})$ 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 defined by

$$\mathbf{F} \equiv \begin{bmatrix} \rho_1 & \rho_2 & \cdots & \rho_{p-1} & \rho_p \\ 1 & 0 & \cdots & 0 & 0 \\ 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & \cdots & 1 & 0 \end{bmatrix}.$$

Using the sequence of autocovariances φ_j obtained in this way, the matrix $\mathbf{\Omega}_i$ can be constructed using $\mathbf{\Omega}_i[j,k] = \varphi_{j-k}$. For example, in the AR(1) case, the jkth element of $\mathbf{\Omega}_i$ is given by $\rho^{|j-k|}/(1-\rho^2)$, i.e.

$$\mathbf{\Omega}_i = \frac{1}{1-\rho^2}\begin{pmatrix} 1 & \rho & \cdots & \rho^{T_i-1} \\ \rho & \ddots & & \vdots \\ \vdots & & \ddots & \rho \\ \rho^{T_i-1} & \cdots & \rho & 1 \end{pmatrix}. \tag{17}$$

Stacking the observations for all n clusters, by analogy with (15), we have

$$\mathbf{z} = \mathbf{X}\beta + \mathbf{Q}\mathbf{g} + \mathbf{W}\mathbf{b} + \varepsilon, \quad \varepsilon \sim \mathcal{N}(0, \mathbf{\Omega}), \tag{18}$$

where $\mathbf{\Omega}$ is a block diagonal matrix given by

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

The fact that the errors are not orthogonal (and $\mathbf{\Omega}$ is not diagonal) requires only minor adjustments to the sampling of β, $\mathbf{z}$, $\mathbf{D}$, τ^2, and $\mathbf{b}$ in Algorithm 1, but the sampling of $\mathbf{g}$ is problematic. The difficulty is that, after accounting for the autocorrelated errors, Step 4 of Algorithm 1 will involve the matrix $\mathbf{G}^{-1} = \left(\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}\right)$ which is not banded any longer (even though $\mathbf{\Omega}^{-1}$ 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 that are 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. There attention is restricted to univariate models for non-clustered 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 $\mathbf{G}^{-1} = \left(\mathbf{K}/\tau^2 + \mathbf{Q}'\mathbf{\Omega}^{-1}\mathbf{Q}\right) = \left(\mathbf{K}/\tau^2 + \mathbf{\Omega}^{-1}\right)$ 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, as repeating values in $\mathbf{s}$ will tend to emerge across clusters. The general case when $\mathbf{s}$ is allowed to be any covariate (not necessarily time) is considered in Smith *et al.* (1998) but their algorithm is $\mathcal{O}\left(N^3\right)$ as it works with the nonbanded precision matrix $\mathbf{G}^{-1}$. Thus, the applicability of that method is only limited to small data sets and is infeasible in panel data settings where the sample size N can easily run into the (tens of) 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, Chapter 6; Chib 1993) works well in parametric models, but will not be a solution here because it is equivalent to premultiplying

the matrices $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Q}$, by the Cholesky decomposition of $\mathbf{\Omega}^{-1}$ and this still leaves $\mathbf{G}^{-1}$ nonbanded.

Below 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 part, and to deal with the correlated part of the errors 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 $\mathbf{\Omega}_i = \mathbf{R}_i + \kappa\mathbf{I}$, 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 $\mathbf{\Omega}_i = \mathbf{C}_i'\mathbf{C}_i + \kappa\mathbf{I}$ and the model can be rewritten as

$$\begin{aligned} \mathbf{z}_i &= \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \mathbf{g}_i + \varepsilon_i \\ &= \mathbf{X}_i\beta + \mathbf{W}_i\mathbf{b}_i + \mathbf{g}_i + \mathbf{C}_i'\mathbf{u}_i + \xi_i, \end{aligned} \tag{19}$$

where $\mathbf{u}_i \sim N(\mathbf{0}, \mathbf{I})$ and $\xi_i \sim N(\mathbf{0}, \kappa\mathbf{I})$ are mutually independent. Stacking the observations in (19) for all n clusters, by analogy with (15) and (18), we have

$$\mathbf{z} = \mathbf{X}\beta + \mathbf{Q}\mathbf{g} + \mathbf{W}\mathbf{b} + \mathbf{C}'\mathbf{u} + \xi, \quad \xi \sim \mathcal{N}(0, \kappa\mathbf{I}), \tag{20}$$

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

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

Since the covariance matrix of ξ 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 $\mathbf{\Omega}_i$ 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. Our proof will be by construction, which is useful in providing intuition about applying the decomposition in practice. The details are formalized below.

Theorem 1 *A symmetric positive definite matrix $\mathbf{\Omega}_i$ 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. Our goal is to write $\mathbf{\Omega}_i = \mathbf{R}_i + \kappa\mathbf{I}$ where we want $\mathbf{R}_i$ to be symmetric and positive definite. The symmetry of $\mathbf{R}_i$ is immediate since $\mathbf{\Omega}_i$ is symmetric and $\kappa\mathbf{I}$ is diagonal. It will then be sufficient to show how to choose $\kappa > 0$ so that $\mathbf{R}_i$ is positive definite. For positive definiteness we require $\mathbf{x}'\mathbf{R}_i\mathbf{x} > 0$ for any $\mathbf{x} \neq \mathbf{0}$. Substituting for $\mathbf{R}_i$, we need $\mathbf{x}'(\mathbf{\Omega}_i - \kappa\mathbf{I})\mathbf{x} > 0$. Since $\mathbf{\Omega}_i$ is symmetric positive definite, it can be written as $\mathbf{\Omega}_i = \mathbf{V}_i\mathbf{\Lambda}_i\mathbf{V}_i'$, where $\mathbf{\Lambda}_i$ is a diagonal matrix containing its (strictly positive and real) eigenvalues $\{\lambda_{ij}\}$, and $\mathbf{V}_i$ is an orthogonal matrix of eigenvectors such that $\mathbf{V}_i'\mathbf{V}_i = \mathbf{V}_i\mathbf{V}_i' = \mathbf{I}$. Therefore, $\mathbf{x}'(\mathbf{\Omega}_i - \kappa\mathbf{I})\mathbf{x} = \mathbf{x}'(\mathbf{V}_i\mathbf{\Lambda}_i\mathbf{V}_i' - \kappa\mathbf{I})\mathbf{x} = \mathbf{x}'(\mathbf{V}_i\mathbf{\Lambda}_i\mathbf{V}_i' - \kappa\mathbf{I}\mathbf{V}_i\mathbf{V}_i')\mathbf{x} = \mathbf{x}'\mathbf{V}_i(\mathbf{\Lambda}_i - \kappa\mathbf{I})\mathbf{V}_i'\mathbf{x}$ and setting $\mathbf{y} = \mathbf{V}_i'\mathbf{x}$, ($\mathbf{y} \neq \mathbf{0}$ because $\mathbf{V}_i$ is nonsingular) we have $\mathbf{y}'(\mathbf{\Lambda}_i - \kappa\mathbf{I})\mathbf{y} = \sum_j (\lambda_{ij} - \kappa) y_j^2$, which is guaranteed to be positive as long as $\min\{\lambda_{ij}\} > \kappa > 0$. ■

As the above 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 very well in our simulations is to set $\kappa = \min\{\lambda_{ij}\}/2$.

Our approach to estimating a model with serial correlation is based on Algorithm 1, the discussion above, and the algorithms in Chib and Greenberg (1994) for sampling of the vector of autoregressive coefficients $\rho = (\rho_1, ..., \rho_p)'$. The prior on ρ is specified as

$$\rho \sim \mathcal{N}(\rho_0, \mathbf{P}_0) I_{S_\rho},$$

where I_{S_ρ} is an indicator of the set S_ρ, defined as the set of ρ that satisfy stationarity. For the sampling of ρ it will be useful to define the following quantities. Let $e_{it} = z_{it} - \mathbf{x}_{it}'\beta - \mathbf{w}_{it}'\mathbf{b}_i - g(s_{it})$, $\mathbf{e}_i = (e_{i,p+1}, ..., e_{i,T_i})'$, $\mathbf{e} = (\mathbf{e}_1', ..., \mathbf{e}_n')'$, and let $\mathbf{E}$ denote the $(N - np) \times p$ matrix whose rows contain p lags of e_{it} ($i = 1, ..., T_i$, $t \geq p+1$), that is $(e_{i,t-1}, ..., e_{i,t-p})$. Finally, let the initial p values of e_{it} in each cluster be given by $\mathbf{e}_{i1} = (e_{i1}, ..., e_{ip})'$ and let $\mathbf{\Omega}_p$ be the $p \times p$ stationary covariance matrix of the AR(p) error process, which is a function of ρ and is constructed in the same way as the $\{\mathbf{\Omega}_i\}$. Sampling for ρ uses the Metropolis-Hastings algorithm (Hastings (1970), Tierney (1994), Chib and Greenberg (1995)).

Algorithm 2 ***Model with State Dependence and AR(p) Serial Correlation***

1. Sample $\{\mathbf{z}_i\}|\mathbf{y}, \mathbf{D}, \mathbf{g}, \beta, \rho$ marginal of $\{\mathbf{b}_i\}$ by drawing for $i \leq n$, $t \leq T_i$

$$z_{it} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 1 \\ \mathcal{TN}_{(-\infty,0]}(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 0 \end{cases}$$

where $\mathcal{TN}_{(a,b)}(\mu_{it}, v_{it})$ is a normal distribution truncated to the interval (a, b) with mean $\mu_{it} = E(z_{it}|\mathbf{z}_{i\setminus t}, \beta, \mathbf{g}_i, \mathbf{V}_i)$ and variance $v_{it} = Var(z_{it}|\mathbf{z}_{i\setminus t}, \beta, \mathbf{g}_i, \mathbf{V}_i)$, with $\mathbf{V}_i = \mathbf{\Omega}_i + \mathbf{W}_i \mathbf{D} \mathbf{W}_i'$, and $\mathbf{\Omega}_i$ determined by ρ as discussed above.

2. Sample $\beta, \{\mathbf{b}_i\}, \{\mathbf{u}_i\} |\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g}_i, \rho$ in one block by drawing

   (a) $\beta|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \mathbf{g}, \rho \sim \mathcal{N}(\hat{\beta}, \mathbf{B})$, where $\hat{\beta} = \mathbf{B}(\mathbf{B}_0^{-1}\beta_0 + \sum_{i=1}^n \mathbf{X}_i' \mathbf{V}_i^{-1} (\mathbf{z}_i - \mathbf{g}_i))$ and $\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^n \mathbf{X}_i' \mathbf{V}_i^{-1} \mathbf{X}_i)^{-1}$;

   (b) $\mathbf{b}_i|\mathbf{y}, \mathbf{D}, \{z_{it}\}, \beta, \mathbf{g}, \rho \sim \mathcal{N}(\hat{\mathbf{b}}_i, \mathbf{B}_i)$ with $\hat{\mathbf{b}}_i = \mathbf{B}_i \mathbf{W}_i' \mathbf{\Omega}_i^{-1} (\mathbf{z}_i - \mathbf{X}_i \beta - \mathbf{g}_i)$ and $\mathbf{B}_i = (\mathbf{D}^{-1} + \mathbf{W}_i' \mathbf{\Omega}_i^{-1} \mathbf{W}_i)^{-1}$ for $i = 1, ..., n$;

   (c) $\mathbf{u}_i|\mathbf{y}, \{z_{it}\}, \{\mathbf{b}_i\}, \beta, \mathbf{g}, \rho \sim \mathcal{N}(\hat{\mathbf{u}}_i, \mathbf{U}_i)$, where $\mathbf{U}_i = \left(\mathbf{I} + \mathbf{C}_i \mathbf{R}_{i2}^{-1} \mathbf{C}_i'\right)^{-1}$ and $\hat{\mathbf{u}}_i = \mathbf{U}_i \mathbf{C}_i \mathbf{R}_{i2}^{-1} (\mathbf{z}_i - \mathbf{X}_i \beta - \mathbf{g}_i - \mathbf{W}_i \mathbf{b}_i)$ for $i = 1, ..., n$.

3. Sample $\mathbf{D}^{-1}|\{\mathbf{b}_i\} \sim \mathcal{W}_p \left\{ r_0 + n, \left(\mathbf{R}_0^{-1} + \sum_{i=1}^n \mathbf{b}_i \mathbf{b}_i'\right)^{-1} \right\}$.

4. Sample $\mathbf{g}|\mathbf{y}, \beta, \{\mathbf{b}_i\}, \tau^2, \{z_{it}\}, \{\mathbf{u}_{i1}\} \sim \mathcal{N}(\hat{\mathbf{g}}, \mathbf{G})$, where $\mathbf{G} = \left(\mathbf{K}/\tau^2 + \mathbf{Q}' \mathbf{R}_2^{-1} \mathbf{Q}\right)^{-1}$ and $\hat{\mathbf{g}} = \mathbf{G}\left(\mathbf{K}\mathbf{g}_0/\tau^2 + \mathbf{Q}' \mathbf{R}_2^{-1} \left(\mathbf{z} - \mathbf{X}\beta - \mathbf{W}\mathbf{b} - \mathbf{C}'\mathbf{u}\right)\right)$. As $\mathbf{G}^{-1}$ is banded, estimation can proceed efficiently as discussed in Remark 1.

5. Sample $\tau^2|\mathbf{g} \sim \mathcal{IG}\left(\frac{\nu_0 + m}{2}, \frac{\delta_0 + (\mathbf{g} - \mathbf{g}_0)' \mathbf{K} (\mathbf{g} - \mathbf{g}_0)}{2}\right)$.

6. Sample $\rho|\mathbf{y}, \mathbf{g}, \beta, \{\mathbf{b}_i\}, \{z_{it}\} \propto \Psi(\rho) \times \mathcal{N}(\hat{\rho}, \mathbf{P}) \times I_{S_\rho}$ where $\hat{\rho} = \mathbf{P}(\mathbf{P}_0 \rho_0 + \mathbf{E}'\mathbf{e})$, $\mathbf{P} = (\mathbf{P}_0 + \mathbf{E}'\mathbf{E})$, and $\Psi(\rho) = |\mathbf{\Omega}_p|^{-n/2} \exp\left(-\frac{1}{2} \sum_{i=1}^n \mathbf{e}_{i1}' \mathbf{\Omega}_p^{-1} \mathbf{e}_{i1}\right)$.

Sampling in Step 6 of Algorithm 2 is through a Metropolis-Hastings step where a proposal draw ρ' is generated from the density $\mathcal{N}(\hat{\rho}, \mathbf{P}) I_{S_\rho}$, and is subsequently accepted as the next sample value with probability $\min\{\Psi(\rho')/\Psi(\rho), 1\}$. If the candidate value ρ' is rejected, the current value ρ is repeated as the next value of the MCMC sample.[4]

[4]The approach for dealing with dependent errors is quite general and can be extended in a straightforward fashion to other correlation structures. One such possibility is the exponentially autocorrelated error sequence considered in Diggle and Hutchinson (1989); however, the method can also handle estimation of general correlation matrices using the algorithms in Chib and Greenberg (1998).

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 in determining the impact of an intervention on one or more of the covariates. Due to the nonlinearity induced by the link function, the state dependence, the serial correlation, the unknown function and the subject-specific random effects, the parameters in the model are difficult to interpret. In addition, the exact form of the conditional probability of response is not available and therefore the effect of a change in a given covariate on that probability is not straightforward to calculate. A change in the covariate affects not only the contemporaneous response but also future values of the dependent variable. Moreover, these effects depend on all other covariates and model parameters. Because of this dependence, we calculate that effect marginalized over the remaining covariates and parameters.

To enhance understanding, suppose the canonical model for a new individual i is given by

$$z_{it} = \mathbf{x}'_{it}\beta + \mathbf{w}'_{it}\mathbf{b}_i + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it},$$

where $\mathbf{x}'_{it} = (\mathbf{x}^{*\prime}_{it}, \mathbf{w}'_{it}\mathbf{A}_i)$, $\beta = (\delta', \gamma')'$, 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 β accordingly, we re-write the above model as

$$z_{it} = x_{1it}\beta_1 + \mathbf{x}'_{2it}\beta_2 + \mathbf{w}'_{it}\mathbf{b}_i + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it}.$$

The average covariate effect can then be analyzed from a predictive perspective applied to this new individual i. Suppose that one thinks of setting x_{1i1} to the value $x^{\dagger}_{1j1}$. For a predictive horizon of $t = 1, 2, ..., T_i$ (where T_i is the smallest of the cluster sizes in the observed data) one is now interested in the distribution of $y_{i1}, y_{i2}, ..., y_{iT_i}$ marginalized over $\{\mathbf{x}_{2it}\}$, $\mathbf{b}_i$, and $\theta = \left(\beta, \phi, \mathbf{g}, \mathbf{D}, \tau^2, \rho\right)$ given the current data $\mathbf{y} = (\mathbf{y}_1, ..., \mathbf{y}_n)$. A practical procedure is to marginalize out the covariates as a Monte Carlo average using their empirical distribution, while θ is integrated out with respect to the posterior distribution $\pi(\theta|\mathbf{y})$. Of course $\mathbf{b}_i$ is independent of $\mathbf{y}$ and hence can be integrated out of the joint distribution of $\{z_{i1}, ..., z_{iT_i}\}$ analytically using the distribution $N(\mathbf{0}, \mathbf{D})$, without recourse to Monte Carlo. Therefore, the goal is to obtain a

sample of draws from the distribution

$$[z_{i1},...,z_{iT_i}|\mathbf{y},\mathbf{y}_{i0},x_{1i1}^{\dagger}] = \int \left[z_{i1},...,z_{iT_i}|\mathbf{y},\mathbf{y}_{i0},x_{1i1}^{\dagger},\{\mathbf{x}_{2it}\},\{\mathbf{w}_{it}\},\{s_{it}\},\theta\right] \pi\left(\{\mathbf{x}_{2it}\},\{\mathbf{w}_{it}\},\{s_{it}\}\right)\pi\left(\theta|\mathbf{y}\right)d\{\mathbf{x}_{it}\}\,d\{\mathbf{w}_{it}\}\,d\{s_{it}\}\,d\theta.$$

In this particular example, there are four possible initial conditions for this subject which means that there are four possible joint distributions for a given value of $x_{1j3}^{\dagger}$.[5] Consider, for example, the case where $\mathbf{y}_{i0}=(0,0)'$. A sample from the above predictive distribution can be obtained by the method of composition applied in the following way. Find all the individuals $N_{00}=\{i:\mathbf{y}_{i0}=(0,0)'\}$. Randomly draw one individual i^* from this set and extract the sequence of covariate values $\{x_{2i^*1},x_{2i^*2},...,x_{2i^*T_i}\}$. Sample a value for θ from the posterior density and sample $\{z_{i1},...,z_{iT_i}\}$ jointly from $\left[z_{i1},...,z_{iT_i}|\mathbf{y},\mathbf{y}_{i0},x_{1i1}^{\dagger},\{\mathbf{x}_{2i^*t}\},\{\mathbf{w}_{i^*t}\},\{s_{i^*t}\},\theta\right]$, 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 $(y_{i1},...,y_{iT_i})$. Repeat this analysis for a different value of x_{1i1}, say $x_{1i1}^{\ddagger}$, and compute the predictive mass function as above. The difference in pointwise probabilities gives the effect of x_1 as it is changed from $x_{1i1}^{\dagger}$ to $x_{1i1}^{\ddagger}$. Finally, repeat these steps for the other three possible initial conditions. The predictive horizon can be extended further into the future, but at the cost of making potentially strong assumptions about the covariates.

The above approach can similarly be applied to other elements of $\mathbf{x}_{it}$, as well as to elements of $\mathbf{w}_{it}$. Quite importantly, it can be applied to determining the effect of the nonparametric component $g(s)$ because the error bands that are usually reported in the estimation of $g(\cdot)$ may not provide sufficient information to make probabilistic statements about differences such as $g\left(s^{\dagger}\right)-g\left(s^{\ddagger}\right)$, in cases when such statements could be meaningful. In addition, the above approach can be done conditionally upon, instead of marginally of, certain variables (e.g. gender, race) that might determine a subsample of interest.

In conclusion, we note that the entire approach in this section can be contrasted with the usual textbook treatment, where in determining the effect of a given covariate, the remaining covariates and parameters are held fixed at their mean values. The drawback of the latter

[5]Since we know the mixing distribution for the initial conditions, we can always produce the joint distribution marginal of the initial conditions, but going in the opposite direction and decomposing the latter distribution into its mixture components is not always possible.

approach is that it ignores the uncertainty in the parameters and the covariates, and will produce misleading results due to the non-linearity of the link function.

5 Model Comparison

A central issue in the analysis of statistical data is model formulation, since 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 versus a semiparametric model), or the distributional assumptions. In general, given the data $\mathbf{y} = (y_1, ..., y_n)$, interest centers upon a collection of models $\{\mathcal{M}_1, ..., \mathcal{M}_L\}$ representing competing hypotheses about the data. Each model $\mathcal{M}_l$ is characterized by a model-specific parameter vector θ_l and sampling density $f(\mathbf{y}|\mathcal{M}_l, \theta_l)$. Bayesian model selection proceeds by comparison of 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

$$\frac{\Pr(\mathcal{M}_i|\mathbf{y})}{\Pr(\mathcal{M}_j|\mathbf{y})} = \frac{\Pr(\mathcal{M}_i)}{\Pr(\mathcal{M}_j)} \times \frac{m(\mathbf{y}|\mathcal{M}_i)}{m(\mathbf{y}|\mathcal{M}_j)} \tag{21}$$

where

$$m(\mathbf{y}|\mathcal{M}_l) = \int f(\mathbf{y}|\mathcal{M}_l,\ \theta_l)\pi_l(\theta_l|\mathcal{M}_l)d\theta_l \tag{22}$$

is the marginal likelihood of $\mathcal{M}_l$. The first fraction on the right hand side of (21) is known as the prior odds and the second as the Bayes factor.

To date in the semiparametric function context, model comparisons have been based on criteria such as the AIC and BIC (e.g. Shively *et al.* 1999, Wood *et al.* 2002, DiMatteo *et al.* 2001, Hansen and Kooperberg 2002), but their computation is infeasible in our setting because the likelihood is intractable. Here we take up the question of calculating the marginal likelihood of our semiparametric model.

The direct evaluation of the integral in (22) is generally infeasible. Chib (1995) provided a method based on the recognition that the marginal likelihood can be re-expressed as

$$m(\mathbf{y}|\mathcal{M}_l) = \frac{f(\mathbf{y}|\mathcal{M}_l, \theta_l)\pi(\theta_l|\mathcal{M}_l)}{\pi(\theta_l|\mathbf{y}, \mathcal{M}_l)}\,. \tag{23}$$

This is an identity that holds for any point θ_l. The calculation of the marginal likelihood is hence reduced to finding an estimate of posterior ordinate $\pi(\theta_l^*|\mathbf{y}, \mathcal{M}_l)$ at a single point θ_l^*. Suppose

that the parameter space is split into B conveniently specified blocks, so that $\theta = (\theta_1, ..., \theta_B)$. Let $\psi_i = (\theta_1, ..., \theta_i)$ denote the blocks up to i and $\psi^{i+1} = (\theta_{i+1}, ..., \theta_B)$ denote the blocks beyond i, and suppress the model index for notational convenience. Then, by the law of total probability we have

$$\pi\left(\theta_1^*, ..., \theta_B^* | \mathbf{y}\right) = \prod_{i=1}^{B} \pi\left(\theta_i^* | \mathbf{y}, \theta_1^*, ..., \theta_{i-1}^*\right) = \prod_{i=1}^{B} \pi\left(\theta_i^* | \mathbf{y}, \psi_{i-1}^*\right), \tag{24}$$

and each of the ordinates $\pi(\theta_i^* | \mathbf{y}, \psi_{i-1}^*)$ can be estimated as

$$\pi\left(\theta_i^* | \mathbf{y}, \psi_{i-1}^*\right) \approx G^{-1} \sum_{g=1}^{G} \pi\left(\theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1,(g)}\right),$$

where $\psi^{i,(g)} \sim \pi\left(\psi^i | \mathbf{y}, \psi_{i-1}^*\right)$, $g = 1, ..., G$, come from a reduced run for $1 < i < B$, and sampling is only over ψ^i, with the blocks ψ_{i-1}^* being held fixed (see Chib (1995)). The ordinate $\pi\left(\theta_1^* | y\right)$ for the first block of parameters θ_1 is estimated with draws $\theta \sim \pi\left(\theta | \mathbf{y}\right)$ from the main MCMC run, while the ordinate $\pi\left(\theta_B^* | \mathbf{y}, \psi_{B-1}^*\right)$ is available directly.

A version of the above method is also available when one or more of the full conditional densities is not of a standard form and sampling requires the Metropolis-Hastings algorithm. Chib and Jeliazkov (2001) use the local reversibility of the M-H Markov chain to show that

$$\pi(\theta_i^* | \mathbf{y}, \psi_{i-1}^*) = \frac{E_1\left\{\alpha\left(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}\right) q\left(\theta_i, \theta_i^* | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}\right)\right\}}{E_2\left\{\alpha\left(\theta_i^*, \theta_i | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}\right)\right\}}, \tag{25}$$

where E_1 is the expectation with respect to conditional posterior $\pi(\psi^i | \mathbf{y}, \psi_{i-1}^*)$ and E_2 that with respect to the conditional product measure $\pi(\psi^{i+1} | y, \psi_i^*) q(\theta_i^*,\ \theta_i | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1})$, where $q(\theta, \theta' | y)$ denotes the candidate generating density of the M-H chain for moving from the current value θ to a proposed value θ', and

$$\alpha(\theta_i, \theta_i' | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}) = \min\left\{1, \frac{f\left(y | \theta', \psi_{i-1}^*,\ \psi^{i+1}\right) \pi\left(\theta', \psi_{i-1}^*, \psi^{i+1}\right) q\left(\theta', \theta | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}\right)}{f\left(\mathbf{y} | \theta, \psi_{i-1}^*, \psi^{i+1}\right) \pi\left(\theta, \psi_{i-1}^*, \psi^{i+1}\right) q\left(\ \theta, \theta' | \mathbf{y}, \psi_{i-1}^*, \psi^{i+1}\right)}\right\}$$

denotes the M-H probability of move from θ to θ'. Each of these expectations can be computed from the output of appropriate reduced runs.

We note that while it is true that the identity (23) can also be written as $m(\mathbf{y} | \mathcal{M}_l) = f(\mathbf{y} | \mathcal{M}_l, \theta_l, \mathbf{z}_l) \pi(\theta_l, \mathbf{z}_l | \mathcal{M}_l) / \pi(\theta_l, \mathbf{z}_l | \mathbf{y}, \mathcal{M}_l)$ when latent variables $\mathbf{z}_l$ are present, that form of the identity is not very useful, because the dimension of $\mathbf{z}_l$ may be very large. We therefore generally integrate out any such parameters before applying (23). Furthermore, we emphasize that the

comparison of dynamic models with different numbers of lags should be based on an equal data sample in order to be meaningful (this is indicated by conditioning on the same data $\mathbf{y}$ in (21)).

In our specific implementation, we decompose the posterior ordinate, marginalized over $\{\mathbf{z}_i\}$ and $\{\mathbf{b}_i\}$, as

$$\pi(\mathbf{D}^*, \tau^{2*}|\mathbf{y})\pi\left(\beta^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}\right)\pi\left(\rho^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*},\ \beta^*\right)\pi\left(\mathbf{g}^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \beta^*, \rho^*\right)$$

and estimate the ordinate of $\mathbf{g}$ last because that tends to improve the efficiency of the ordinate estimation. It may also be noted that the ordinate $\pi\left(\mathbf{D}^*, \tau^{2*}|\mathbf{y}\right)$ can be estimated jointly because Algorithms 1 and 2 reveal that conditional on $\{\mathbf{b}_i\}$ and $\mathbf{g}$, the full conditional densities of $\mathbf{D}$ and τ^2 are independent. This observation saves the need to do an additional reduced run. Second, in Algorithm 2 the proposal density $q\left(\rho, \rho'|\mathbf{y}, \cdot\right) = q\left(\rho'|\mathbf{y}, \cdot\right)$ is a truncated normal density that has an unknown normalizing constant except for the AR(1) case. Specifically, over the region of stationarity S_ρ we have $q\left(\rho|\mathbf{y}, \cdot\right) = h\left(\rho|\mathbf{y}, \cdot\right) / \int_{s_\rho} h\left(\rho|\mathbf{y}, \cdot\right) \mathrm{d}\rho$, where $h\left(\rho|\mathbf{y}, \cdot\right)$ is an unrestricted normal density for ρ. The problem is very similar to the one discussed in Chib and Jeliazkov (2004a), and we adapt their solution here. Following that approach, it can be shown that the reversibility condition used by Chib and Jeliazkov (2001) to obtain (25) can be re-written in terms of $q\left(\rho|\mathbf{y}, \cdot\right)$ whose unknown normalizing constant on both sides will cancel, so that upon integration, $\pi\left(\rho^*|\mathbf{y}, \mathbf{D}^*, \tau^{2*}, \beta^*\right)$ can be estimated by

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

where $\psi^*_{i-1} = \left(\mathbf{D}^*, \tau^{2*}, \beta^*\right)$, $\psi^{i+1} = \left(\mathbf{g}, \{\mathbf{z}_i\}, \{\mathbf{b}_i\}\right)$, E_1 is the expectation with respect to conditional posterior $\pi(\rho, \psi^{i+1}|\mathbf{y}, \psi^*_{i-1})$, and E_2 that with respect to the conditional product measure $\pi(\psi^{i+1}|y, \psi^*_i)h(\rho|\mathbf{y}, \psi^*_{i-1}, \psi^{i+1})$. It is also worth noting that because $h\left(\rho'|\mathbf{y}, \cdot\right)$ does not depend on the current value for ρ in the sampler, estimation of the denominator quantity is done with draws that are available from the same run in which the numerator is estimated. Further analysis of the model comparison method is taken up in Chib and Jeliazkov (2004b), where marginal likelihoods are estimated for different models, and the correctness of the estimates is verified for cases in which answers are available by alternative estimation methods.

The implementation of (23), marginal of $\{\mathbf{z}_i\}$ and $\{\mathbf{b}_i\}$, requires the likelihood ordinate $f\left(\mathbf{y}|\mathbf{D}^*, \beta^*, \tau^{2*}, \mathbf{g}^*, \rho^*\right)$. To obtain this ordinate, we use the Geweke, Hajivassiliou, and Keane

(GHK) method, which provides estimates of the likelihood contributions (14) at the values $\left(\mathbf{D}^*, \beta^*, \tau^{2*}, \mathbf{g}^*\right)$. The method is based on writing $\mathbf{V}_i = \mathbf{L}\mathbf{L}'$, where $\mathbf{L}$ is a lower triangular Cholesky factorization, and making a change of variable from $\mathbf{z}_i$ to ε_i where $\mathbf{z}_i = \mathbf{X}_i\beta + \mathbf{g_i} + \mathbf{L}\varepsilon_i$. Then

$$\begin{aligned}\Pr(\mathbf{y}_i|\mathbf{y}_{i0}, \beta^*, \mathbf{g}_i^*, \mathbf{D}^*) &= \int_{B_{iT_i}} \cdots \int_{B_{i1}} N_{T_i}(\mathbf{z}_i|\mathbf{X}_i\beta^* + \mathbf{g}_\mathbf{i}^*, \mathbf{V}_i)d\mathbf{z}_i \\ &= \int_{c_{iT_i}^*}^{d_{iT_i}^*} \cdots \int_{c_{i1}^*}^{d_{i1}^*} N_{T_i}(\mathbf{u}|\mathbf{0}, \mathbf{I})d\mathbf{u},\end{aligned}$$

where

$$c_{it}^* = \frac{c_{it} - x_{it}'\beta^* - g_{it}^* - \sum_{k=1}^{t-1} l_{tk}\varepsilon_{ik}}{l_{tt}}, \quad d_{it}^* = \frac{d_{it} - x_{it}'\beta^* - g_{it}^* - \sum_{k=1}^{t-1} l_{tk}\varepsilon_{ik}}{l_{tt}},$$

and c_{it} and d_{it} denote the lower an upper limits of integration of B_{it} respectively. The integral is then estimated by recursive Monte Carlo simulation, and the likelihood ordinate is obtained as the product of the estimates of the individual likelihood contributions. In the example, we use 10000 Monte Carlo iterations.

6 Simulation Study

The key aspect of our implementation is that it relies on a fully Bayesian, finite sample methodology for the analysis of the model in Section 2. This is enabled by our use of proper priors for the parameters and the unknown function $g(\cdot)$, and may be contrasted with previous studies (Silverman 1985, Wood and Kohn 1998, Hastie and Tibshirani 2000, Fahrmeir and Lang 2001), where partially improper priors are used. In our simulation study we calculate mean squared errors for the estimates of the unknown function, which are reported for several designs. The posterior mean estimates $E\{g(\mathbf{v})|\mathbf{y}\}$, are found from MCMC runs of length 5000 following burn-ins of 1000 draws. A second goal for this study is to demonstrate the performance of the MCMC estimation algorithm by reporting the autocorrelations and the inefficiency factors for the sampled parameters under alternative model specifications and sample sizes. We find that the overall performance of the MCMC algorithm improves with larger sample sizes (either with larger number of clusters n or with larger cluster sizes $\{T_i\}$), and that the random effects are simulated better when the increase in sample size comes as a result of increasing the cluster sizes $\{T_i\}$.

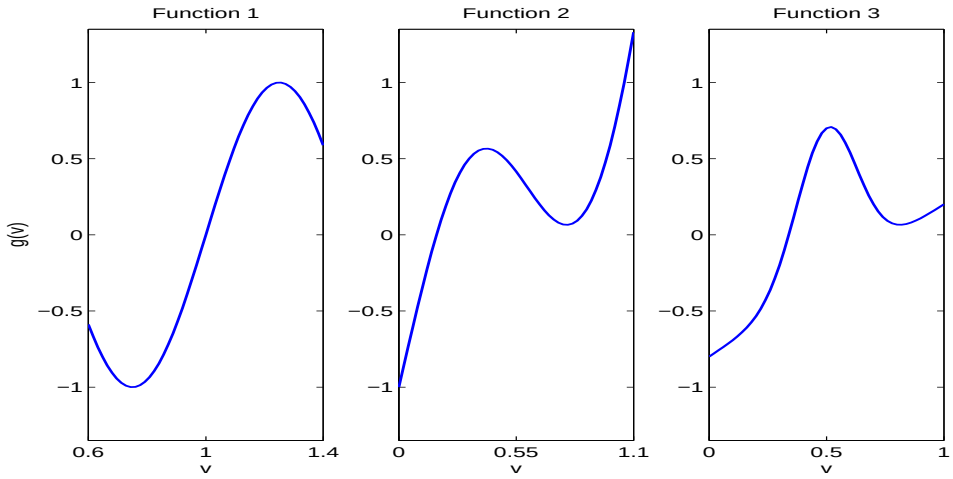

Figure 1: The three true functions in the simulation study.

Data is simulated from the model in (1) and (3), using 1, 2, and 3 lags, a single fixed effect covariate $\mathbf{X}$, and 1 or 2 individual effect covariates $\mathbf{W}$ (including a random intercept). $\mathbf{X}$ and $\mathbf{W}$ contain independent standard normal random variables, and we use $\delta = 1$, $\gamma = \mathbf{1}$, $\phi = 0.5 * \mathbf{1}$, and $\mathbf{D} = 0.2 * \mathbf{I}$. We generate panels with 250, 500, and 1000 clusters, and with 10 time periods, using only the last 7 for estimation ($T_i = 7$, $i = 1, ..., n$), since our largest models contain 3 lags (the initial conditions are treated as given and are generated randomly). We consider three functional specifications for the function g:

1. $g(s) = \sin(2\pi s)$, for $s \in [0.6, 1.4]$;
2. $g(s) = -1 + s + 1.6s^2 + \sin(5s)$, for $s \in [0, 1.1]$;
3. $g(s) = -0.8 + s + \exp\left\{-30(s - 0.5)^2\right\}$, for $s \in [0, 1]$.

The three functions are plotted in Figure 1. Each of them is evaluated on a regular grid of $m = 51$ points. We have chosen these functions to capture a range of possible specifications – for example, the first function achieves its extrema in the interior of its domain, while the second does so at the endpoints of the domain; the third function has a minimum at the end, and a maximum in the interior, of its domain. In addition, the first function is symmetric, while the other two are asymmetric. We gauge the performance of the method in fitting the above functions using mean squared error

$$MSE = \frac{1}{m} \sum_{j=1}^{m} \{\hat{g}(v_j) - g(v_j)\}^2 . \tag{26}$$

Clusters	Lags	Random effects	Average Mean Squared Errors g_1	g_2	g_3
n = 250	$J = 1$	$q = 1$	0.01804 (0.01022)	0.01421 (0.00998)	0.01751 (0.01135)
		$q = 2$	0.01039 (0.00537)	0.02463 (0.01231)	0.01694 (0.00712)
	$J = 2$	$q = 1$	0.02102 (0.01331)	0.01406 (0.00711)	0.01717 (0.00659)
		$q = 2$	0.02329 (0.02497)	0.03128 (0.03427)	0.03189 (0.02481)
	$J = 3$	$q = 1$	0.02528 (0.03216)	0.02411 (0.02436)	0.02678 (0.01621)
		$q = 2$	0.03063 (0.02459)	0.02080 (0.01756)	0.02232 (0.02274)
n = 500	$J = 1$	$q = 1$	0.00609 (0.00259)	0.00591 (0.00414)	0.00755 (0.00394)
		$q = 2$	0.00914 (0.00574)	0.00815 (0.00361)	0.01232 (0.05837)
	$J = 2$	$q = 1$	0.00816 (0.00521)	0.00755 (0.00393)	0.01225 (0.00623)
		$q = 2$	0.00902 (0.00653)	0.00890 (0.00576)	0.00885 (0.00930)
	$J = 3$	$q = 1$	0.01427 (0.00780)	0.01590 (0.00978)	0.01913 (0.00717)
		$q = 2$	0.01461 (0.01038)	0.01716 (0.01384)	0.01896 (0.02248)
n = 1000	$J = 1$	$q = 1$	0.00332 (0.00194)	0.00627 (0.00432)	0.00404 (0.00167)
		$q = 2$	0.00573 (0.00326)	0.00482 (0.00241)	0.00566 (0.00317)
	$J = 2$	$q = 1$	0.00324 (0.00171)	0.00387 (0.00175)	0.00562 (0.00317)
		$q = 2$	0.00683 (0.00677)	0.00634 (0.00226)	0.00662 (0.00529)
	$J = 3$	$q = 1$	0.00727 (0.00501)	0.00622 (0.00302)	0.00551 (0.00238)
		$q = 2$	0.00838 (0.00418)	0.00943 (0.00761)	0.01096 (0.00776)

Table 1: Average mean squared errors based on 10 samples, with estimated standard errors in parentheses.

The average MSE, together with the standard errors based on 10 data samples, is reported in Table 1 for various specifications. It is important to keep in mind that the goal of Table 1 is not to illustrate the best possible fit for every possible situation, because that fit will depend on the assumed priors and the specific data sample. The goal of Table 1 is rather to illustrate the relative performance of the method under alternative specifications and also to show that given the assumed priors, as the sample size grows the function will be estimated arbitrarily well. In all cases, we have used comparable mildly informative priors, amongst which of particular importance is the prior on τ^2, which determines the appropriate degree of smoothness (more on this below).

From Table 1 we see that as the sample size grows, in all cases the functions are estimated more and more precisely, as expected. Also, in line with conventional wisdom, the general trend seems to be that fitting models with fewer parameters for a given sample size 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, it leads to increasing the number of parameters in the model, and second, it affects the proportion of ones among the responses (since all elements in ϕ are positive). It is well known that the degree of asymmetry in the proportion of the responses affects the estimation precision. For our one-lag models, the proportion of ones is between 0.62 and 0.67 across the three functional specifications, for the two-lag models that proportion is between 0.67 and 0.72, and for the three lag models, it is between 0.71 and 0.76. As Table 1 shows, however, the method recovers the true functions well, despite this asymmetry. As an illustration of the technique, in Figure 2 we show three particular nonparametric function fits for $n = 500$, $J = 2$, and $q = 1$.

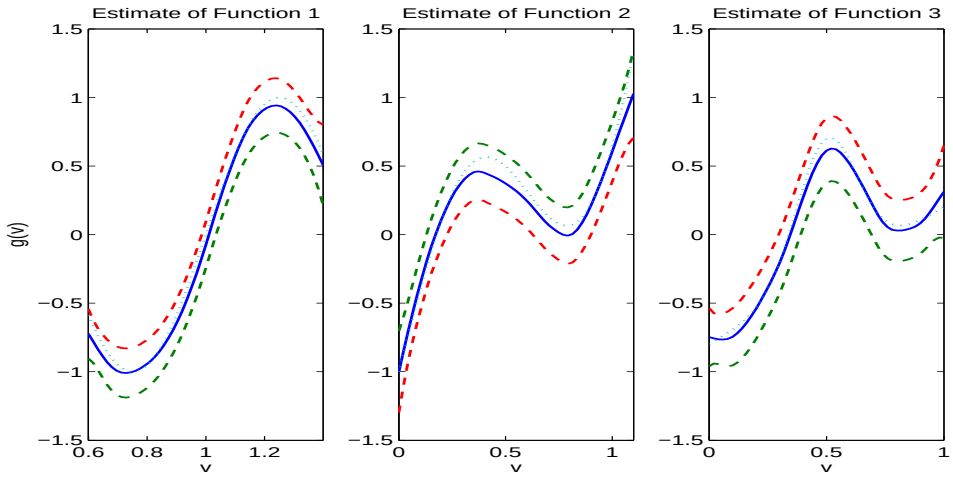

Figure 2: Simulation Study. Three examples of estimated functions (solid lines), true functions (dotted lines), and 95% confidence bands (dashed lines).

Considering the model as a whole, it should be apparent that we have a latent variable model with three important variance structures, two of which involve parameters to be estimated (τ^2 and $\mathbf{D}$) and the other is fixed for identification purposes (the error distribution has variance 1 in the probit case). Because of this, it is important to be aware that the relative informativeness or noninformativeness of the priors for these variances should be viewed in the context of the other variance priors, and not in isolation. The usual motivation for considering this interdependence is that 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 $g(\cdot)$ (with this trade-off being the focal point of the penalized likelihood approach to nonparametric regression). Similarly, the variance of the errors and the variance of the random effects $\mathbf{D}$ determine a balance between intra- and inter-

cluster variation. While in large samples the effects of the assumed priors on the parameter estimates is small (vanishing asymptotically), in small samples informative priors do matter. Figure 3 illustrates this point by using two somewhat exaggeratedly different informative priors on τ^2. In one case the prior on τ^2 is such that $E\left(\tau^2\right) = 0.5$ and $\mathrm{SD}\left(\tau^2\right) = 0.1$; in the second case $E\left(\tau^2\right) = 0.001$ and $\mathrm{SD}\left(\tau^2\right) = 0.001$. The figure illustrates that the first prior leads to a function which is more wiggly as it curves to interpolate the data more closely, while the second prior leads to oversmoothing. When the sample size is increased from $n = 250$ to $n = 1000$, the difference in the function estimates becomes much smaller, but the different degree of smoothness is clearly visible in the graphs.

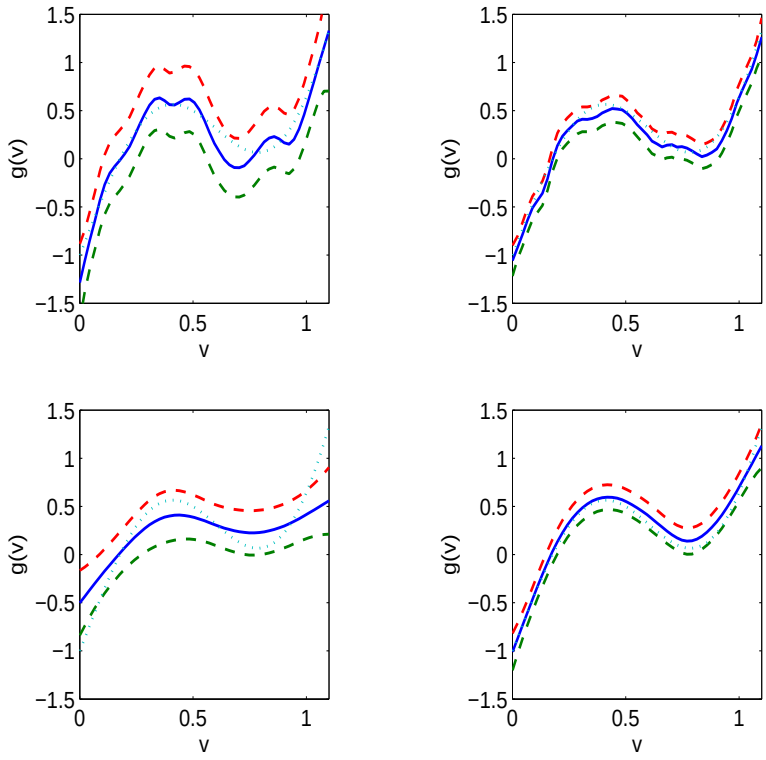

Figure 3: Effect of τ^2 on the estimates of $\mathbf{g}$ (solid lines) for two sample sizes: $n = 250$ in the first column, and $n = 1000$ in the second. Since τ^2 is large in the first row, the estimated function is less smooth; a small τ^2 in the second row leads to oversmoothing. The confidence bands (dashed lines) are tighter in the second column because of the larger sample size.

An example of the performance of the MCMC sampler for the problem with $n = 500$ (with $T_i = 7$) is illustrated in Figure 4, which shows an example of histograms and kernel-smoothed marginal posterior densities for the parameters together with the corresponding autocorrelations from the MCMC sampler. The linear effects, together with τ^2 appear to be estimated well and

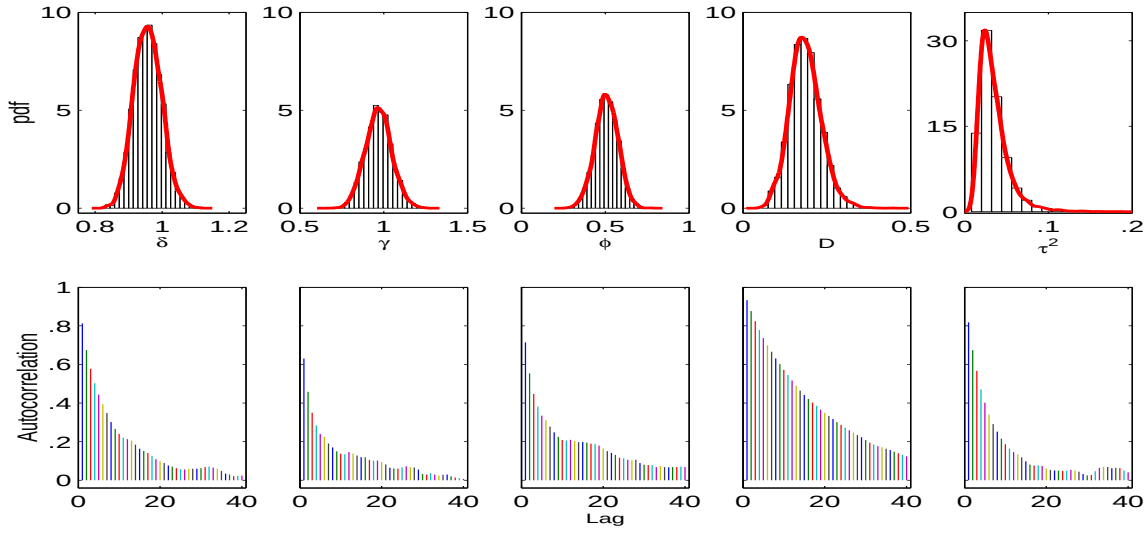

Figure 4: Posterior samples and autocorrelations for the parameters of a semiparametric model with one fixed effect, one random effect, and one lag ($T_i = 7$, $i = 1, ..., n$).

the sample is characterized by low autocorrelations. While it can be seen that $\mathbf{D}$ is estimated well, its higher autocorrelation indicates that its mixing is slower than that of the remaining parameters, and because of this longer Markov chain runs may be needed in order to describe the marginal posterior density of $\mathbf{D}$ more accurately. The slower mixing occurs because $\mathbf{D}$ is a parameter at the second level of the modeling hierarchy and depends on the data only indirectly through $\{\mathbf{b}_i\}$ (i.e. given $\{\mathbf{b}_i\}$, $\mathbf{D}$ does not depend on $\{\mathbf{z}_i\}$ and $\mathbf{y}$). Since $\{\mathbf{b}_i\}$ are not well identified in smaller clusters, when only a few observations are available to identify the cluster-specific effects, and because learning about $\mathbf{D}$ occurs from the inter-cluster variation of $\{\mathbf{b}_i\}$, $\mathbf{D}$ also suffers from weak identification when cluster sizes are small. To measure the efficiency of the MCMC parameter sampling scheme we use the measures $[1 + 2\sum_{k=1}^{\infty} \rho_k(l)]$, where $\rho_k(l)$ is the sample autocorrelation at lag l for the kth parameter in the sampling with the summation truncated according to (say) the Parzen window. The latter quantity is called the *inefficiency factor* or the *autocorrelation time* and 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 shows the inefficiency factors corresponding to the parameters for the same model as above, but now with different cluster sizes (7, 12, and 17 observations per cluster). In this setup the larger cluster size serves to identify $\{\mathbf{b}_i\}$ better, allowing for inter-cluster variation to be captured more precisely. In line with the arguments

	Inefficiency Factors		
Parameter	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	13.683	10.183	12.553
γ	9.617	6.155	7.247
ϕ	12.111	8.310	9.031
$\mathbf{D}$	24.002	13.276	9.413
τ^2	10.654	5.260	8.740

Table 2: Examples of estimated inefficiency factors (autocorrelation times) for the parameters of the model with one lag, one random effect, and one fixed effect for $n = 500$.

above, Table 2 shows that the inefficiency factor for $\mathbf{D}$ drops considerably (the other inefficiency factors stay within a similar range). The improvement in the sampling of $\mathbf{D}$ is also easily seen from a comparison of Figures 4 and 5, with the latter summarizing the MCMC output used for the third column of Table 2, when $T_i = 17$.

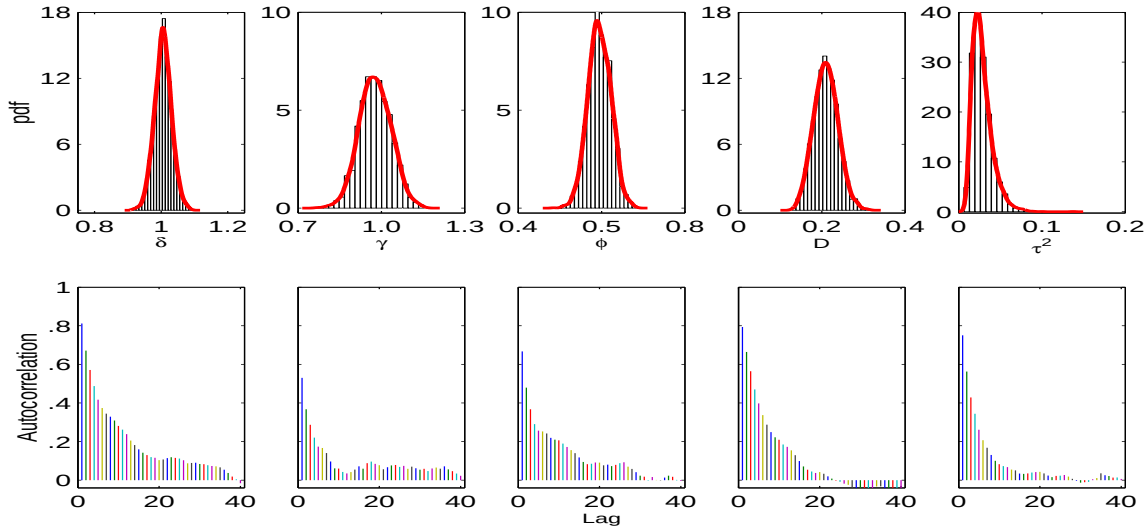

Figure 5: Posterior samples and autocorrelations for the parameters of a semiparametric model with one fixed effect, one random effect, and one lag ($T_i = 17$, $i = 1, ..., n$).

Similarly to the cases discussed in Table 2, we present results for the inefficiency factors for a model with two lags and two random effects in Table 3. Since now not one, but two random effects are estimated from the limited observations in each cluster, the elements of $\mathbf{D}$ 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 $\mathbf{D}$ drop noticeably.

Finally, in Table 4, we report results from experiments involving a model with serially corre-

	Inefficiency Factors								
	n=250			n=500			n=1000		
Param	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	21.156	18.093	18.882	19.885	16.072	14.721	20.788	22.462	20.650
γ_1	12.192	12.195	11.511	14.725	12.995	12.815	12.164	13.567	12.336
γ_2	11.399	8.483	9.607	15.638	8.836	7.325	14.640	10.073	9.564
γ_3	9.059	10.364	8.414	11.750	10.477	9.191	12.592	12.903	8.436
ϕ_1	9.288	9.121	7.996	7.016	10.962	11.065	14.363	10.229	10.806
ϕ_2	8.385	7.599	10.140	7.692	7.521	8.480	9.147	11.141	10.009
$\mathbf{D}_{11}$	27.008	22.457	18.865	31.560	23.884	18.722	34.094	25.949	18.075
$\mathbf{D}_{12}$	29.728	23.157	18.156	29.884	25.814	14.908	26.313	20.249	18.379
$\mathbf{D}_{22}$	26.243	26.372	17.627	34.718	27.870	18.092	34.556	24.668	21.799
τ^2	10.580	12.453	10.166	14.742	9.474	7.802	8.015	6.051	6.015

Table 3: Examples of estimated inefficiency factors (autocorrelation times) for the parameters of the model with two lags, two random effects and one fixed effect.

lated errors. Table 4 contains the inefficiency factors for three cluster sizes and two values of ρ. The parameters ρ and $\mathbf{D}$ are sampled well in all cases, but it is interesting to see that when ρ is positive, both ρ and $\mathbf{D}$ 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, $\mathbf{D}$ is identified better, so both ρ and $\mathbf{D}$ are estimated better. This does not appear to be a problem when $\rho < 0$, because then the two correlation structures implied by ρ and $\mathbf{D}$ are quite different. For the samplers and values of ρ being considered, the M-H acceptance rate in the sampling of ρ was in the range of $(0.87, 0.98)$.

	Inefficiency Factors ($\rho = -0.5$)			Inefficiency Factors ($\rho = 0.5$)		
Parameter	$T_i = 7$	$T_i = 12$	$T_i = 17$	$T_i = 7$	$T_i = 12$	$T_i = 17$
δ	19.059	17.430	20.190	20.748	14.064	17.715
γ	9.340	7.299	7.864	11.421	11.303	12.046
ϕ	23.047	25.626	18.235	17.001	20.239	20.108
$\mathbf{D}$	21.475	16.214	12.751	39.468	28.021	15.191
τ^2	9.680	7.076	6.010	11.837	7.965	7.970
ρ	20.501	21.345	20.276	32.398	27.122	26.447

Table 4: Examples of estimated inefficiency factors (autocorrelation times) for the parameters of the model with one lag, one random effect, one fixed effect, and AR(1) serial correlation for $n = 500$.

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 identified better. Most noticeably, the sampling of $\mathbf{D}$ benefits strongly from the availability of 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 1545 married women in the age range of 17-66. The data set, based on Hyslop (1999), 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 9 covariates, which are presented 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. The data set is obtained from the Panel Study of Income Dynamics (PSID), which is available on the web at http://www.isr.umich.edu/src/psid/ and contains information on education, employment, income, family composition, and many other variables of economic interest. Similar data have been analyzed by Chib and Greenberg (1998), who estimated multivariate probit models using MCMC methods, by Avery, Hansen, and Hotz (1983) using the method of moments, and by Hyslop (1999) who fits dynamic probit models by maximum simulated likelihood estimation, and compares the estimates to those from linear probability models and static probit models. Covariates similar to those in Table 5 are also common in empirical models of the intensive (hours), in addition to the extensive (participation), margin of female labor supply (e.g. Nakamura and Nakamura (1994), Shaw (1994), Heckman (1993), Mroz (1987)).

A key feature of our application is that the effect of age on the conditional probability of working is modeled nonparametrically. Nonlinearities arise due to changes in trade-offs and tastes for work over a woman's life cycle, age-related changes in health (both her own, and of her close relatives), the fact that age is indicative of the expected timing of events (graduation from school or college, marriage, planning for children, etc.), and because a woman's age may be revealing of her social values and education type (cohort effect), her experience as a homemaker and in the market (productivity effect), and the types of jobs available to her. Previous studies

Variable	Explanation	Mean	SD
WORK	wife's labor force status (1=working, 0=not working)	0.7097	0.4539
INT	an intercept term (a column of ones)		
AGE	the woman's age in years	36.0262	9.7737
RACE	1 if black, 0 otherwise	0.1974	0.3981
EDU	attained education (in years) at time of survey	12.4858	2.1105
CH2	number of children aged 0-2 in that year	0.2655	0.4981
CH5	number of children aged 3-5 in that year	0.3120	0.5329
CH13	number of children aged 6-13 in that year	0.6763	0.8851
CH17	number of children aged 14-17 in that year	0.2950	0.6064
INC	total annual labor income of head of household	31.7931	22.6417

Table 5: Variables in the women's labor force participation application. The dependent variable is the woman's labor force status, and the remaining variables are explanatory variables. The summary statistics are based on the entire sample of observations. INC (in thousands of dollars) is measured as nominal earnings adjusted by the consumer price index (base year = 1987).

have attempted to capture some of this nonlinearity by including polynomials in age (Hyslop (1999)), or by considering separate age groups (e.g. Shaw (1994), Blau (1998), Nakamura and Nakamura (1994)). It is well known that the results will be contingent upon the particular choice of age groups, and that parametric models offer only limited flexibility and affect the shape of the regression function globally, rather than locally – for these reasons nonparametric modeling may be preferable. We note, however, that even if one prefers to use parametric models for the ultimate purpose of explanation or prediction, a semiparametric model offers an important exploratory step toward final model determination.

A second important aspect of the current application is that state dependence is incorporated through two lags of the dependent variable. The second lag leads to improved model performance and a higher marginal likelihood than for a model with only one lag. This finding is quite sensible in light of the existence of multiple sources of state dependence, whose effects cannot *a priori* be restricted to single-lag specifications. Such sources of state dependence include human capital accumulation (e.g. Heckman (1981)), search costs of finding a new job (e.g. Hyslop (1999), Eckstein and Wolpin (1990)), costs of solving additional practical problems (child care needs, transportation, relocation of housework, resolution of scheduling conflicts) which would have already been solved by employed women (Nakamura and Nakamura (1994)), and intertemporal nonseparability of preferences for leisure (Hotz, Kydland, and Sedlacek (1988)). From an applied perspective, this implies that reliance upon single-lag models without allowing for more elaborate

forms of state dependence may have adverse consequences, as controlling for dynamic factors is crucial for eliciting other substantive issues of interest, such as heterogeneity and covariate effects.

Third, a principal finding of this analysis is that in addition to variation in the intercept, other important sources of heterogeneity emerge in considering the impact of small children on female labor supply. The issue was studied by Angrist and Evans (1998), who were interested in the effect that a woman's schooling or her husband's earnings may have on the labor supply consequences of childbearing. Interest in this problem stems from a number of theoretical models of home production and the value of a housewife's time, which suggest a dependence on her education and her or her husband's income (Gronau (1973a, 1973b, 1977), Angrist and Evans (1996)). As an empirical issue, Angrist and Evans (1998, p. 469) argue that "...because mothers' education and husbands' wages are correlated, it is not clear whether a set of estimates that condition on husbands' earnings and a set of estimates that condition on mothers' education are capturing distinct phenomena" (in our sample, $Corr(EDU, INC) = 0.30$ and $Corr(EDU, \ln(INC)) = 0.32$). We address this and other issues of model determination further below, and in anticipation of those results we present our baseline specification, where, in addition to a random intercept, the effects of the covariates CH2 and CH5 are specified as random and are allowed to depend on husbands' earnings and the initial conditions.

Summarizing the above discussion, our main model ($\mathcal{M}_1$) is given by:

$$\Pr(y_{it} = 1|\theta, \beta_i) = 1\left\{\mathbf{x}_{it}^{*\prime}\delta + \mathbf{w}_{it}^{\prime}\beta_i + g\left(s_{it}\right) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it} > 0\right\},$$

$$\beta_i = \mathbf{A}_i\gamma + \mathbf{b}_i, \quad \mathbf{b}_i \sim N_3(0, \mathbf{D}),$$

where $y_{it} = WORK_{it}$, $\theta = (\delta, \mathbf{g}, \phi_1, \phi_2, \gamma, \mathbf{D})$, $\mathbf{x}_{it}^{*\prime} = (RACE, EDU_{it}, \ln(INC_{it}))$, $s_{it} = AGE_{it}$, $\mathbf{w}_{it}^{\prime} = (1, CH2_{it}, CH5_{it})$, and

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

Because of its features, model $\mathcal{M}_1$ can be contrasted with previous studies of female labor supply which largely rely on parametric models with at most one lag, under the assumption that variation in the intercept is the only source of heterogeneity.

Parameter	Covariate	Mean	SD	Median	Lower	Upper	Ineff
δ	$RACE$	0.170	0.080	0.169	0.014	0.329	7.012
	EDU	0.087	0.015	0.086	0.057	0.117	23.189
	$\ln(INC)$	-0.190	0.048	-0.189	-0.286	-0.098	16.484
γ	$\bar{y}_{i0}$	1.371	0.173	1.365	1.047	1.724	27.802
	$CH2$	0.142	0.312	0.144	-0.479	0.747	5.414
	$(CH2)(\bar{y}_{i0})$	-0.245	0.161	-0.248	-0.556	0.077	19.356
	$(CH2)\left(\overline{\ln(INC_i)}\right)$	-0.135	0.093	-0.135	-0.318	0.046	6.230
	$CH5$	0.868	0.273	0.867	0.339	1.416	8.358
	$(CH5)(\bar{y}_{i0})$	-0.351	0.127	-0.350	-0.606	-0.103	14.530
	$(CH5)\left(\overline{\ln(INC_i)}\right)$	-0.221	0.081	-0.221	-0.380	-0.063	8.139
ϕ	$y_{i,t-1}$	1.213	0.071	1.213	1.072	1.348	15.863
	$y_{i,t-2}$	0.445	0.071	0.445	0.308	0.581	11.470
$vech(\mathbf{D})$		0.540	0.129	0.528	0.319	0.828	38.481
		-0.043	0.096	-0.043	-0.243	0.133	45.999
		0.137	0.071	0.119	0.046	0.319	45.617
		-0.151	0.085	-0.138	-0.347	-0.019	43.454
		0.017	0.049	0.011	-0.066	0.136	45.551
		0.158	0.086	0.135	0.047	0.366	46.355
τ^2		0.017	0.006	0.016	0.009	0.030	5.473

Table 6: Parameter estimates for model $\mathcal{M}_1$. The table also reports 95% confidence intervals and inefficiency factors from 15000 MCMC iterations.

The parameter estimates for $\mathcal{M}_1$ are presented in Table 6. Interpretation of the estimates in Table 6 is complicated by the nonlinearity of the problem and the interactions between the variables. For example, the income and child variables are important determinants of female labor supply but they enter the model in a way that makes it difficult to disentangle and evaluate their effects. For this reason, we present the average effects for certain changes in these covariates in Figure 6. More specifically, the figure presents the average effects of three hypothetical scenarios: first, doubling of the husband's earnings, second, the effect of an additional birth in period 1 (i.e. having an additional child aged 0-2 in periods 1-3, who grows and changes categories to become a child aged 3-5 in periods 4 and 5), and third, the effect of an additional child aged 3-5 in periods 1-3. The figure presents results for sets of individuals distinguished by their initial conditions, namely $N_{mn} = \left\{i : \mathbf{y}_{i0} = (m,n)'\right\}$, as well as overall results for the sample.[6]

[6]The results for the overall effect are produced by averaging the results for the subsets $\{N_{mn}\}$ with respect to their proportion relative to the entire data set, using the observed occurrences: $\#\{N_{00}\} = 337$, $\#\{N_{01}\} = 98$, $\#\{N_{10}\} = 127$, $\#\{N_{11}\} = 983$.

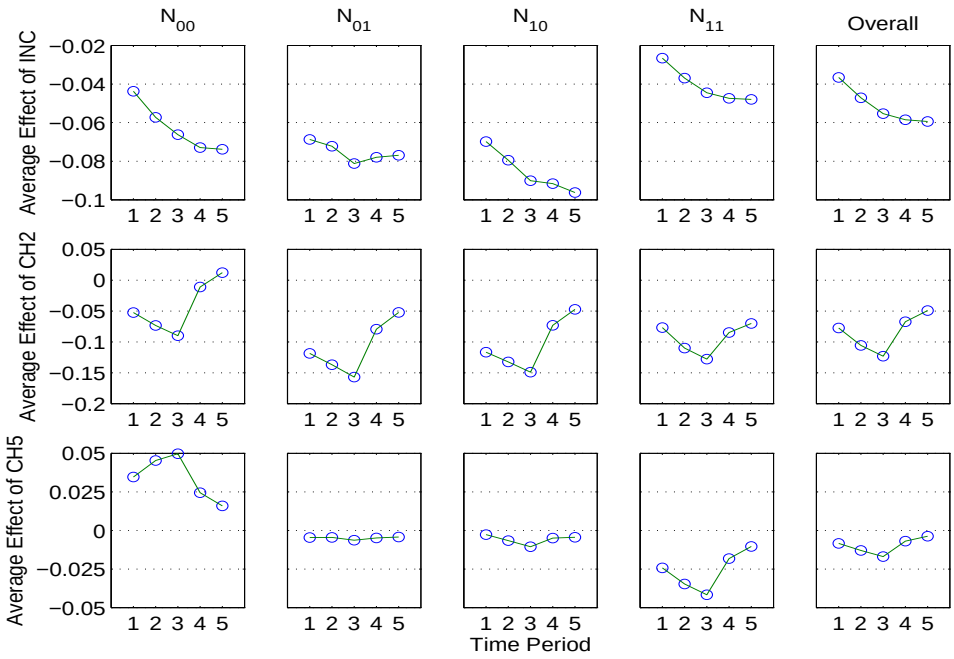

Figure 6: The average effect of doubling a husband's permanent income (first row), having an additional child aged 0-2 in periods 1-3, and aged 3-5 in periods 4 and 5 (second row), or an additional child aged 3-5 in periods 1-3 (third row).

From Table 6 and Figure 6, we see that conditional on the covariates, black women appear more likely to work, and that, *ceteris paribus*, women who work are more likely to be better educated, or have husbands with low earnings (although quite large changes in earnings are necessary to produce economically significant changes in participation). Quite importantly, the results in Table 6 indicate strong state dependence on the first two lags, so that women who have worked in the previous two years are much more likely to work this year and vice versa. Figure 6 shows that there is a negative overall effect of pre-school children on labor supply, which is noticeably stronger for children aged 0-2 than for children aged 3-5. The results also show that after controlling for state dependence and the remaining covariates, the impact of husband's earnings on the effect of children on a woman's participation is negative, agreeing with theoretical predictions. This finding is even more significant when contrasted with the result in Angrist and Evans (1998), who found that their estimates contradicted the theoretical prediction that the labor supply of more educated women responds more to the presence of children. There are many differences between the two models and the estimation techniques, and one such difference is that in $\mathcal{M}_1$ husband's earnings, rather than wife's education, are

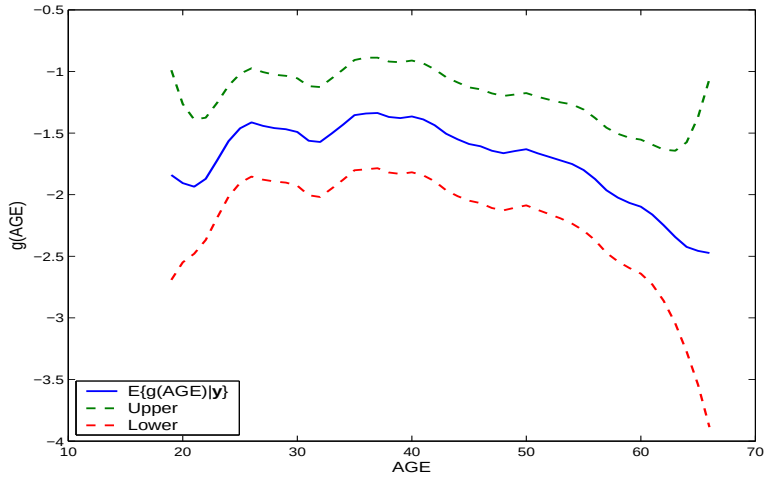

Figure 7: The effect of age on the probability of working.

allowed to be correlated with the child effects – our conclusions did not change when we fit a model where education was correlated with the child effects, however, the marginal likelihood for that model was lower than that of $\mathcal{M}_1$ as discussed at the end of this section.

There is strong correlation between the initial conditions and the random effects in this model. In particular, the positive correlation between the random intercept and the initial conditions and the negative correlation between $\bar{y}_{i0}$ and the effects of CH2 and CH5 is in agreement with a human capital theory, and also with the fact that the initial observations are indicative of a woman's tastes. More precisely, these correlations, as captured by the parameters γ, are consistent with the explanation that in equilibrium, the marginal returns of an hour spent at home are higher for higher productivity women, and those will tend to be ones who have worked in the initial period. Increasing the number of children will then increase the time spent at home and reduce the probability of working, but more so for higher productivity women in the presence of increasing returns to scale in child-rearing (Angrist and Evans (1996, 1998)).

The estimate of the nonparametric function $g\left(AGE\right)$ is shown in Figure 7. From Figure 7 we see that the impact of age is characterized by interesting nonlinearities in women's 20's and 30's, and that a woman's propensity to participate in the labor force drops off as she approaches retirement age. In particular, we see that the probability of employment grows quickly in the early twenties, as women graduate from college and begin work, then decreases somewhat in the late twenties and early thirties only to increase again in the mid-thirties and early forties.

The drop-off in the function after the mid-forties is consistent with a cohort effect (women who were over 40 in the early 1980's were raised and educated at a time when the expectation was that they will be housewives), as well as the effects of age-related changes in health (both her own, and of her older relatives). The peculiar behavior of the function around age 30 represents an interesting question for future research, for which we do not have an explanation at present. In several alternative models, including a random intercept model, the dip in the probability of working around age 31 was somewhat more pronounced, making an approximation by low-order polynomials in age less revealing of the impact of age. Overall, the features of the regression relationship presented in Figure 7 were quite stable across several alternative semiparametric specifications, including single lag models or models with different covariates, which mostly resulted in vertical shifts of $g\,(AGE)$. The log marginal likelihood of $\mathcal{M}_1$, the baseline model discussed above, was estimated to be -2563.826.

A number of alternative model specifications were considered. Issues such as variable selection, lag determination, and correlation between the unobserved effects and covariates, are handled as model selection problems by computing the marginal likelihoods and Bayes factors of competing models. In the interest of clarity and completeness, the more important model determination issues are revisited in Table 7. We begin with the problem of variable selection. Many previous articles considered INC as a covariate, but because of the strong degree of skewness of the income distribution, INC had little explanatory power. Angrist and Evans (1998), considered using $\ln(INC)$ rather than INC, and using this covariate transformation here, perhaps not surprisingly, resulted in decisively higher marginal likelihoods (all differences were over 20 on the natural log scale) across the alternatives we considered. In addition, in agreement with the general view in labor economics, the results from this model support the proposition that a woman's decision to work is affected mainly by pre-school children, and not by older children. More specifically, models including CH13 and CH17 (either as fixed or random effects) had lower marginal likelihoods than models without these covariates – see models $\mathcal{M}_2$ and $\mathcal{M}_3$ in Table 7.

Turning attention to the heterogeneity in the individual specific effects, we see from Table 7 that model $\mathcal{M}_4$, where the conditional means of the child status effects are allowed to be correlated with $\overline{EDU_i}$ instead of with $\overline{\ln(INC_i)}$, does not perform as well as $\mathcal{M}_1$ (using both

Model	Fixed Eff.	Random Eff.	Non-zero elements in $\mathbf{A}_i$	ln(Marg. Lik.)
Baseline model:				
$\mathcal{M}_1$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$(\bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{\ln(INC_i)};$ $1, \bar{y}_{i0}, \overline{\ln(INC_i)})$	-2563.826
Models with CH13 and CH17:				
$\mathcal{M}_2$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2},$ $CH13_{it}, CH17_{it}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	-2574.326
$\mathcal{M}_3$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}, CH13_{it}, CH17_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	-2598.028
Models with alternative heterogeneity assumptions:				
$\mathcal{M}_4$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$(\bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{EDU_i};$ $1, \bar{y}_{i0}, \overline{EDU_i})$	-2572.240
$\mathcal{M}_5$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$(\bar{y}_{i0}; 1, \bar{y}_{i0}; 1, \bar{y}_{i0})$	-2568.647
$\mathcal{M}_6$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$(\bar{y}_{i0}, \overline{\ln(INC_i)};$ $1, \bar{y}_{i0}, \overline{\ln(INC_i)};$ $1, \bar{y}_{i0}, \overline{\ln(INC_i)})$	-2589.771
$\mathcal{M}_7$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	1	$(\bar{y}_{i0})$	-2580.102
Parametric models:				
$\mathcal{M}_8$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2},$ AGE_{it}	$\mathbf{w}_{it}$	$(1, \bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{\ln(INC_i)};$ $1, \bar{y}_{i0}, \overline{\ln(INC_i)})$	-2581.156
$\mathcal{M}_9$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2},$ AGE_{it}, AGE^2_{it}	$\mathbf{w}_{it}$	$(1, \bar{y}_{i0}; 1, \bar{y}_{i0}, \overline{\ln(INC_i)};$ $1, \bar{y}_{i0}, \overline{\ln(INC_i)})$	-2574.563
Single-lag model:				
$\mathcal{M}_{10}$	$\mathbf{x}^*_{it}, y_{i,t-1}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	-2610.709
Model with dependent errors:				
$\mathcal{M}_{11}$	$\mathbf{x}^*_{it}, y_{i,t-1}, y_{i,t-2}$	$\mathbf{w}_{it}$	$\mathbf{A}_i$ as in $\mathcal{M}_1$	-2579.551

Table 7: Alternative models in the women's labor force participation application. In this table, we use $\mathbf{x}^*_{it} = (RACE_{it}, EDU_{it}, \ln(INC_{it}))'$ and $\mathbf{w}_{it} = (1, CH2_{it}, CH5_{it})'$ and except for the parametric models, the effect of age is modeled nonparametrically. Only the non-zero elements of $\mathbf{A}_i$ are presented, with commas separating the columns in a given row, and semi-colons separating rows. The log marginal likelihoods are estimated from MCMC runs of length 15000.

$\overline{EDU_i}$ and $\overline{\ln(INC_i)}$ in $\mathbf{A}_i$ performed even worse, and is not reported). Husbands' earnings appear to have richer information content than wives' education in this particular application, despite the fact that the two are closely correlated. Two additional competing specifications are presented in $\mathcal{M}_5$ and $\mathcal{M}_6$. Model $\mathcal{M}_5$ allows the individual effects to be correlated with the initial conditions but not with any covariates, while $\mathcal{M}_6$ allows all random effects, including the intercept, to depend on $\overline{\ln(INC_i)}$ and the initial conditions. Although one of the specifications is more parsimonious, while the other is less parsimonious than $\mathcal{M}_1$, both have lower marginal likelihoods than $\mathcal{M}_1$, illustrating that Bayes factors support the inclusion of relevant covariates but penalize overparameterization. Most importantly, the table shows that a "traditional" specification with a single unobserved effect (a random intercept) did worse than $\mathcal{M}_1$ by a large

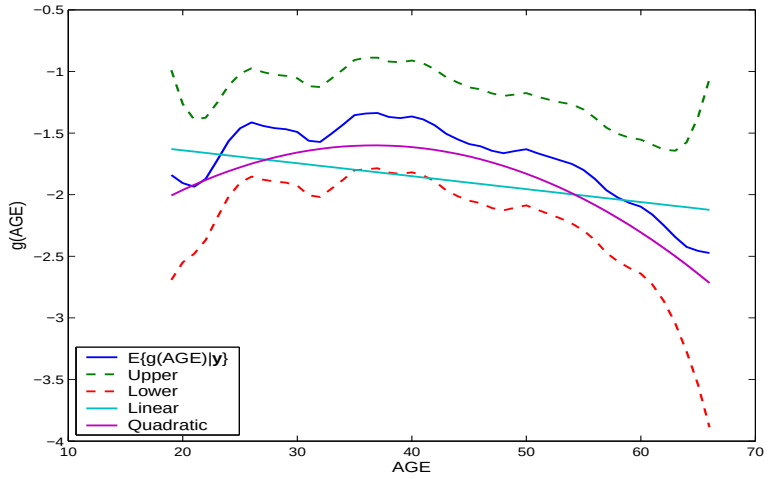

Figure 8: Parametric estimates, resulting from a linear and a quadratic specification in age, versus the nonparametric estimate of the function $g(AGE)$ reported previously.

margin as can be seen from $\mathcal{M}_7$.

The index function of model $\mathcal{M}_8$ is linear in age, but the model is otherwise similar to $\mathcal{M}_1$ with 2 lags and 3 random effects (observe that $\mathbf{A}_i$ is not restricted for identification purposes since now $g(\cdot)$ is not general). That model produced a negative coefficient estimate for age of -0.0105 with 95% credibility region given by $(-0.018, -0.003)$. The estimate from this parametric model can be deceiving, because it overlooks the drastic increase in the probability of working in women's early twenties. A more flexible parametric model is $\mathcal{M}_9$, which uses a quadratic in age. The estimates from the two parametric models are plotted against the nonparametric estimate from $\mathcal{M}_1$ in Figure 8. From the figure we see that the estimates are generally very close, but that even the more complex parametric model still underestimates the strong increase in $g(AGE)$ in women's early twenties.

A final important point relates to the state dependence in the model. Models with a single lag resulted in marginal likelihoods which were lower by over 30 on the natural log scale than models with two lags – the single-lag version ($\mathcal{M}_{10}$) of our baseline model had a marginal likelihood which was lower than that of $\mathcal{M}_1$ by about 47 on the natural log scale. In this example, higher order state dependence turns out to be extremely significant in another way as well. One peculiarity of Hyslop's (1999) results is the presence of statistically significant negative serial correlation in the errors. Hyslop (p. 1288) states that "*[a] suitable interpretation for this is not obvious. One possibility, beyond the scope of this paper, is that the form of state dependence is*

misspecified, and that the AR(1) error component is acting as a fitting parameter in the model. For example, individuals' human capital, which affects their wage offers and depends on their past participation decisions, will imply a more general form of state dependence." Our results from a model with AR(1) serially correlated errors ($\mathcal{M}_{11}$) confirm Hyslop's conjectures entirely. Adding the second lag in the specification (as well as accounting more fully for additional sources of heterogeneity and nonlinearity) leave no serial correlation in the errors. In the sampling of model $\mathcal{M}_{11}$ the value of ρ was estimated to be -0.047 with a 95% confidence interval $(-0.224, 0.113)$. As indicated in Table 7, the marginal likelihood of -2579.551 is significantly lower than that of our baseline model. We take the results from the analysis as a strong warning urging us to be careful when we too quickly accept the ability of single-lag models to properly account for state dependence.

The methodology described in the paper has allowed for the analysis of features of this application such as (i) nonlinearity in the conditional probability of working, (ii) multi-lag state dependence, (iii) serial correlation in the errors, (iv) heterogeneity in the effect of multiple covariates, and (v) correlation between the random effects and the covariates. The techniques developed in Sections 2-4 provide a flexible and conceptually straightforward semiparametric framework for analyzing the above complexities, while guarding against overparameterization by using marginal likelihoods to judge the evidence in the data in favor of particular modeling decisions regarding variable and model selection. The approach is also useful in providing interpretable results in terms of the average covariate effects. The broader implications emerging from our analysis are that the complexity of real-world panel data applications should warrant an extended analysis of the above concerns and that issues of model determination should not be underestimated. Our application shows that more involved modeling can be used to uncover interesting insights and improve the fit, but also that model complexity does not necessarily guarantee better performance – simpler models with fewer covariates or simpler structures often outperform more complex counterparts.

8 Concluding Remarks

This paper has provided a hierarchical Bayes framework for analyzing dynamic binary panel data. Key advantages of the modeling approach that distinguish it from previous work are the semiparametric modeling of the conditional response probability, and the ability to accommodate general state dependence, serial correlation, and multiple unobserved effects. The approach provides a useful mechanism for dealing with uncertainty in function estimation, as well as for describing (rather flexibly) important dependence relations of the unobserved effects on covariates and the initial observations. The techniques rely on latent variable augmentation and a proper Markov process smoothness prior on the unknown function.

The paper has also extended existing econometric methods in order to produce tuned MCMC algorithms for simulation of the posterior distribution, for estimation of the marginal likelihood, and for describing the average covariate effects. Consequently, we can address the problem of model choice and implement a simulation-based approach that enables interpretation of the estimates. The fitting algorithms are computationally efficient, permitting the analysis of large data panels, even in the presence of serially correlated errors. A simulation study shows that the method performs well, and that its performance improves with larger samples. In an application involving a dynamic semiparametric model of women's labor force participation we illustrate that the model and the estimation methods are practical and can help uncover interesting and important features of the data. In particular, the application indicates that models with a single lag and a random intercept may perform inadequately in addressing the complexity of women's intertemporal labor force participation decisions. In comparison, Bayes factors strongly support a model where two lags of the dependent variable enter the probability of working, and where, in addition to a random intercept, the effects of pre-school children on labor supply are unit-specific and correlated with husband's earnings.

One benefit of the model considered above is that it can be easily 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.

References

Albert, J. and S. Chib (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.

Angrist, J. and W. Evans (1996): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," National Bureau of Economic Research (Cambridge, MA) Working Paper 5406.

Angrist, J. and W. Evans (1998): "Children and Their Parents' Labor Supply: Evidence from Exogenous Variation in Family Size," *The American Economic Review*, 88, 450-477.

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

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

Blau, F. (1998): "Trends in the Well-Being of American Women, 1970-1995," *Journal of Economic Literature*, 36, 112-165.

Chamberlain, G. (1984): "Panel Data," in *Handbook of Econometrics*, Vol. II, eds. Z. Griliches and M. Intriligator. Amsterdam: North Holland.

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

Chib, S. (1995): "Marginal Likelihood from the Gibbs Output," *Journal of the American Statistical Association*, 90, 1313-21.

Chib, S. (2001): "Markov Chain Monte Carlo Methods: Computation and Inference," in *Handbook of Econometrics*, Volume 5 (eds J. Heckman and E. Leamer), North Holland, Amsterdam, 3569-3649.

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

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

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

Chib, S. and E. Greenberg (1998): "Analysis of Multivariate Probit Models," *Biometrika*, 85, 2, 347-361.

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

Chib, S. and I. Jeliazkov (2004a): "Accept-Reject Metropolis-Hastings Sampling and Marginal Likelihood Estimation," Technical Report, Washington University.

Chib, S. and I. Jeliazkov (2004b): "Estimation and Model Choice in Nonparametric Additive Regression," Technical Report, Washington University.

Craven, P. and G. Wahba (1979): "Smoothing Noisy Data With Spline Functions," *Numer. Math.*, 31, 377-403.

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

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

DiNardo, J., and J. Tobias (2001): "Nonparametric Density and Regression Estimation," *Journal of Economic Perspectives*, 15, 11-28.

Eckstein, Z. and K. Wolpin (1990): "On the Estimation of Labor Force Participation, Job Search, and Job Matching Models Using Panel Data," in *Advances in the Theory and Measurement of Unemployment*, Y. Weiss and G. Fishelson (eds.), New York: Macmillan.

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

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

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

Gronau, R. (1973a): "The Intrafamily Allocation of Time: The Value of the Housewives' Time," *The American Economic Review*, 63, 634-651.

Gronau, R. (1973b): "The Effect of Children on the Housewife's Value of Time," *The Journal of Political Economy*, 81, S168-S199.

Gronau, R. (1977): "Leisure, Home Production and Work – the Theory of the Allocation of Time Revisited," *The Journal of Political Economy*, 85, 1099-1124.

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

Hansen, M. H., and C. Kooperberg (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: Phillip Allen.

Hastie, T. and R. Tibshirani (2000), "Bayesian Backfitting" (with discussion), *Statistical Science*, 15, 196-223.

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*, S. Rosen (ed.), 91-131. University of Chicago Press.

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

Hsiao, C. (1986): *Analysis of Panel Data.* Cambridge, U.K.: Cambridge University Press.

Honoré, B. and E. Kyriazidou (2000): "Panel Data Discrete Choice Models with Lagged Dependent Variables," *Econometrica*, 68, 839-874.

Hotz, V., Kydland, F., and G. Sedlacek (1988): "Intertemporal Preferences and Labor Supply," *Econometrica*, 56, 335-360.

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

Klaassen, F., and J. Magnus (2001): "Are Points in Tennis Independent and Identically Distributed? Evidence from a Dynamic Binary Panel Data Model," *Journal of the American Statistical Association*, 96, 500-509.

Kohn, R., Ansley, C. F., and D. Tharm (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 D. J. Poirier (2004): "Bayesian Variants of Some Classical Semiparametric Regression Techniques," *Journal of Econometrics*, forthcoming.

Mroz, T. (1987): "The Sensitivity of an Empirical Model of Married Women's Hours of Work to Economic and Statistical Assumptions," *Econometrica*, 55, 765-799.

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

Nakamura, A. and M. Nakamura (1994): "Predicting Female Labor Supply: Effects of Children and Recent Work Experience," *The Journal of Human Resources*, 29, 304-327.

Shaw, K. (1994): "The Persistence of Female Labor Supply: Empirical Evidence and Implications." *The Journal of Human Resources*, 29, 348-378.

Shiller, R. (1973): "A Distributed Lag Estimator Derived From Smoothness Priors," *Econometrica*, 41, 775-788.

Shiller, R. (1984): "Smoothness Priors and Nonlinear Regression," *Journal of the American Statistical Association*, 79, 609-615.

Shively, T. S., R. Kohn, and S. Wood (1999), "Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior" (with discussion), *Journal of the American Statistical Association*, 94, 777-806.

Silverman, B. (1985): "Some Aspects of the Spline Smoothing Approach to Non-parametric Regression Curve Fitting" (with discussion), *Journal of the Royal Statistical Society, B*, 47, 1-52.

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

Tierney, L. (1994), "Markov chains for exploring posterior distributions," (with discussion), *Annals of Statistics*, 22, 1701-1762.

Wooldridge, J. (2000): "A Framework for Estimating Dynamic, Unobserved Effects Panel Data Models With Possible Feedback to Future Explanatory Variables," *Economics Letters*, 68, 245-250.

Wahba, G. (1978): "Improper Priors, Spline Smoothing and the Problem of Guarding Against Model Errors in Regression," *Journal of the Royal Statistical Society, B*, 40, 364-372.

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

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

Yatchew, A. (1998): "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature*, 36, 669-721.