Semiparametric Hierarchical Bayes Analysis of Discrete Panel Data with State Dependence

Siddhartha Chib*
Washington University, St. Louis, MO 63130, USA.

Ivan Jeliazkov†
Washington University, St. Louis, MO 63130, USA.

June, 2002

Abstract

In this paper we consider the analysis of semiparametric models for binary panel data with state dependence. A hierarchical modeling approach is used for dealing with the initial conditions problem, for addressing heterogeneity, and for incorporating correlation between the covariates and the random effects. We consider a semiparametric model in which a Markov process prior is used to model an unknown regression function. Estimation is done by computationally efficient Markov chain Monte Carlo methods that exploit the underlying latent structure of binary data models, following Albert and Chib (1993). Simulation results suggest that the method performs well. In addition to estimation, we address the problem of model selection which is a key concern when dealing with a multitude of possible 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 applied to modeling of 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; Markov process priors.

1 INTRODUCTION

Dynamic binary response models are of particular interest in applications where past occurrence of an event may enter the probability of another occurrence in some genuine physiological,
behavioral, or structural way by altering preferences, relevant constraints, or trade-offs. Traditional examples include the analysis of labor union participation, investment, health status and illness recurrence, fertility decisions, accident occurrence, and (un)employment, among others. Panel data offer a unique opportunity to differentiate between “true state dependence” (where lagged dependent variables are relevant to the outcome) and the phenomenon of “spurious state dependence”, where temporal pseudodependence emerges because of heterogeneity – if subjects possess different temporally persistent unobserved propensities to experience the event, history may simply serve as a proxy (Heckman (1978, 1981), Hsiao (1986, Chapter 7.4)).

In important recent work, Honoré and Kyriazidou (2000) and Hyslop (1999) have addressed some of the inferential challenges in the analysis of binary panel data with state dependence. Honoré and Kyriazidou (2000) analyze the unobserved effects panel data logit model with covariates and one or two lags of the dependent variable under general assumptions about the distribution of the unobserved effect. They also discuss how the analysis can be conducted in a semiparametric model which relaxes the logit assumption. Honoré and Kyriazidou (2000) note that their approach relies on an identification argument that does not allow time dummies and that the analysis requires four or more observations per individual. Moreover, only certain sequences of outcomes can be used for estimation. While the dynamic probit model in Hyslop (1999) avoids some of the above difficulties with estimation, it is less flexible than the model in Honoré and Kyriazidou (2000) on other margins, since it is cast into a more rigid parametric framework. Furthermore, we note that in both models heterogeneity is reflected only in the intercept, and that generalizations to an arbitrary number of time periods and more than one or two lagged dependent variables are not straightforward. The problem of model selection is not addressed.

This paper makes the following main contributions. First, central to the analysis is the fact that the index function of the conditional probability of response is partially linear; in other words, the effect of a given covariate $s$ in the index function is modeled nonparametrically, along
the lines of Wahba (1978), Shiller (1984), Wood and Kohn (1998) and Fahrmeir and Lang (2001). Second, in contrast to previous approaches, the semiparametric approach taken here allows for finite sample inferences and model comparisons based on marginal likelihoods and Bayes factors. Third, the current approach explicitly accounts for the possibility that more than one lagged dependent variable may affect the probability of response, conditional on the covariates and the unobserved effects. This conditional probability is allowed to depend not just on a single unobserved effect but on several, so that heterogeneity is not restricted to the intercept. Fourth, we model the dependence of the unobserved effects on the initial observations and the covariates – we assume, in the spirit of Mundlak (1978), Chamberlain (1984), and Wooldridge (2000), that the conditional distribution of the random effects is Gaussian with mean value that depends on the initial observations and the covariates. We discuss, towards the end of the paper, how the Gaussian assumption can be relaxed. Finally, we develop a framework for calculating and describing the average effect of a given covariate \( x \) on the probability of response, both contemporaneously and over time (due to the dynamic structure of the model). This is a useful summary because the nonlinearity of the link function in binary models makes it difficult to evaluate the effect of a covariate, since that effect will depend on the remaining covariates and parameters in the model.

The rest of the paper is organized as follows. In Section 2 we present the model of interest. Section 3 is devoted to the fitting method. The fitting method is relatively inexpensive and relies on modern Markov chain Monte Carlo techniques (Chib and Greenberg (1995, 1996), Chib (2001)) that are tailored to the model at hand. In this section we also show how the average effects of the covariates on the probability of response are calculated. Section 4 is concerned with the comparison of the general model with alternatives based on the marginal likelihood and Bayes factor criteria. Section 5 presents a detailed simulation study of the performance of the estimation and model comparison methods. In Section 6 we consider model extensions and related fitting techniques while Section 7 presents an application of the model and estimation
technique to the intertemporal labor force participation decisions of a panel of married women. Some brief concluding remarks are given in the final section.

\section{THE MODEL}

Let \( y_{it} \) be the binary response variable of interest, where the indices \( i \) and \( t \) (\( i = 1, \ldots, n \), \( t = 1, \ldots, T_i \)) 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 \( x_{it}^* \) and \( w_{it} \), and nonparametrically on the (scalar) covariate \( s_{it} \) in the form

\begin{equation}
\Pr(y_{it} = 1|\delta, \beta_i, g, \phi_1, \ldots, \phi_J) = F\left(x_{it}^*\delta + w_{it}\beta_i + g(s_{it}) + \phi_1 y_{it-1} + \ldots + \phi_J y_{it-J}\right),
\end{equation}

where \( F : \mathbb{R} \to [0, 1] \) is a known link function, \( \delta \) and \( \beta_i \) are vectors of fixed and unit-specific effects, respectively, \( \phi_1, \ldots, \phi_J \) are lag coefficients, and \( g(\cdot) \) is an unknown function (the dependence on the covariates and the lags of the dependent variable is suppressed in the conditioning set in (1)). The vectors \( x_{it}^* \) and \( w_{it} \) contain two disjoint sets of covariates, and because \( g(\cdot) \) is not restricted, \( x_{it}^* \) does not include an intercept or the covariate \( s_{it} \), although those may be modeled as unit-specific (as elements of \( w_{it} \)) under appropriate identification assumptions – we provide details in Section 2.1 below. We parameterize the link function \( F \) as the standard normal cdf, resulting in a probit model, but note that this Gaussian specification readily extends to a Student-t link (Albert and Chib (1993)), or to some known mixture of normals link function, including the logit specification (Wood and Kohn (1998)).

An important feature of the model in (1) is that its dynamic structure accommodates multiple lags of the dependent variable and multiple random effects (not restricted to the intercept only). Another interesting aspect of the model is that, since \( F \) is nondecreasing, the covariates which enter the model linearly affect \( \Pr(y_{it} = 1|\cdot) \) monotonically. This, however, is not necessarily the case with the covariate \( s \), since \( g(\cdot) \) need not be monotonic over its domain. This advantage of our specification over models with linear link function arguments holds even if the link function is estimated flexibly, because flexible link function estimation still retains monotonicity.
The model in (1) has not been specified before in the analysis of dynamic binary panel data; partially linear models have been shown to be effective and have been widely used in continuous data settings (Ahn and Powell (1993), Hausman and Newey (1995), Subramanian and Deaton (1996), Chevalier and Ellison (1997), Heckman et al. (1998), Yatchew (2000)). For an overview of partially linear models and further references, see the review articles by DiNardo and Tobias (2001) and Yatchew (1998). In the following two subsections, we use a hierarchical approach to complete the specification of the model.

2.1 Hierarchical Modeling of the Unobserved Effects

The possibility to account for dependence amongst the effects and the explanatory variables is an important advantage of panel data. We thus consider the conditional distribution of the unobserved effects given the initial conditions and the covariates, aiming to address some underlying issues of correlation that frequently arise in practical applications. For instance, in cases where the effect of a medical treatment is heterogeneous across the individuals in a panel, one may wish to examine the hypotheses that the effect of the treatment is correlated with various covariates, such as income, age, race, or gender. In another application, the hysteresis hypothesis maintains that increasing short-term employment may have long-term effects because it stimulates the acquisition of valuable work experience. Therefore, one may expect that the initial observations on the employment status of an individual may be correlated with an unobserved effect underlying the probability of future employment. Moreover, those initial observations may also influence the effectiveness of various training programs aimed at improving job skills, because those skills may depend on the worker’s experience.

To understand the modeling approach, it is helpful to consider the implications of the model in (1) for the observations in the $i$th cluster. Let $\mathbf{y}_{i0} \equiv (y_{i,-J+1}, ..., y_{i0})$ be the $J$-vector of initial observations for subject $i$ and denote the remaining $T_i$ observations in the cluster by $\mathbf{y}_i \equiv (y_{i1}, ..., y_{iT_i})'$. Also let $\mathbf{y}_{i,-j}$ represent the $T_i$-vector of $j$th order lags of $\mathbf{y}_i$, $j = 1, ..., J,$
and let $\phi = (\phi_1, ..., \phi_J)'$ represent the vector of lag coefficients whose $j$th entry is the coefficient corresponding to $y_{i,-j}$. Finally, define the covariate matrices $X_i^* = (x_{i1}', ..., x_{iT_i}')'$, $W_i = (w_{i1}', ..., w_{iT_i}')'$, and $L_i = (y_{i,-1}, ..., y_{i,-J})$, and let $g_i = (g(s_{i1}), ..., g(s_{iT_i}))'$ be the vector of functional evaluations of $g(\cdot)$ at the points in the vector $s_i = (s_{i1}, ..., s_{iT_i})'$. Then the conditional probability of a particular vector $y_i$, representing the sequence of occurrences for the $i$th subject, is given by

$$\Pr(y_i|y_{i0}, \delta, \phi, \beta_i, g_i) = \prod_{t=1}^{T_i} \Pr(y_{it}|y_{i,-J+1}, ..., y_{i,t-1}, \delta, \phi, \beta_i, g_i)$$

$$= \prod_{t=1}^{T_i} \int_{B_{it}} N(z_{it}|x_{it}'\delta + w_{it}'\beta_i + g(s_{it}) + \phi_1 y_{i,t-1} + ... + \phi_J y_{i,-J}, 1)dz_{it}$$

$$= \int_{B_{it}} \cdots \int_{B_{iT_i}} N_T(z_i|X_i^*\delta + W_i\beta_i + g_i + L_i\phi, I)dz_i,$$  \hspace{1cm} (2)

where $B_{it}$ is the interval $(0, \infty)$ if $y_{it} = 1$, or the interval $(-\infty, 0]$ if $y_{it} = 0$, and $z_i = (z_{i1}, ..., z_{iT_i})'$ is a $T_i \times 1$ vector of latent variables. In the preceding, the first line follows from the law of total probability, the second line follows from assuming a $J$-lag probit specification, and the third line follows from the conditional independence of the $\{z_{it}\}$.

The presence of the initial observations $y_{i0}$ in the conditioning set for each subject, coupled with the possibility for correlation between the covariates and the individual effects $\beta_i$, requires a specification for the conditional distribution of the unobserved effects. Here we assume, using the suggestions in Mundlak (1978), Chamberlain (1984), and Wooldridge (2000), that the distribution of $\beta_i$, conditional on the initial observations and the covariates, is Gaussian with mean value that depends on the initial observations and the covariates. In particular, the $q$-vector of unobserved effects $\beta_i$ is modeled as

$$\beta_i|y_{i0}, X_i^*, W_i, s_i, \gamma, D \sim N_q(A_i \gamma, D), \hspace{1cm} i = 1, ..., n,$$  \hspace{1cm} (3)

where the dependence on the covariates and the initial conditions is made explicit in the conditioning set, and is incorporated through the matrix $A_i$, whose composition is discussed next. An extension beyond the Gaussian case will be discussed in Section 6.
The conditional distribution (3) is equivalent to assuming

\[ \beta_i = A_i \gamma + b_i, \quad b_i \sim N_q(0, D), \quad i = 1, ..., n, \]  

(4)

where the matrix \( A_i \) in (3) and (4) can be defined quite flexibly, given the specifics of the problem at hand. In the simplest case where \( W_i \) does not include an intercept and \( \beta_i \) is independent of the covariates, a parsimonious way of modeling the dependence of \( \beta_i \) on \( y_{i0} \) is to let \( A_i \) be a \( q \times 2q \) matrix given by \( A_i = I \otimes (1, \tilde{y}_{i0}) \), where \( \tilde{y}_{i0} = (1/J) \sum_{j=1}^{J} y_{ij} \) is the mean of the entries in \( y_{i0} \). More generally, the matrix \( A_i \) 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. If \( \tilde{r}_{ij} \) \( (j = 1, ..., q) \) denotes the vector of such covariate means, the general \( A_i \) may be written as

\[ A_i = \begin{pmatrix} 1 & \tilde{y}_{i0} & \tilde{r}_{i1}' & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \ddots & \ddots & \vdots \\ 0 & \cdots & 0 & 1 & \tilde{y}_{i0} & \tilde{r}_{iq}' \end{pmatrix} \]  

(5)

where the dimension of \( \tilde{r}_{ij} \) depends on how many column averages of \( X_i^*, W_i, \) and \( s_i \) were allowed to influence the respective random effect.\(^1\) In constructing the matrix \( A_i \), a less parsimonious, but potentially useful specification may call for the inclusion in \( A_i \) of the entire vector of lags \( y_{i0} \), rather than just the mean \( \tilde{y}_{i0} \), with similar modeling also applicable to the covariates which are correlated with the unobserved effects. This approach is more general in that it allows for differential correlations across time periods, however one has to be aware that it quickly leads to overparameterization, especially with more than one unobserved effect or moderately large number of lags \( J \) or number of periods \( T_i \).

Finally, in order to allow for the most common specification in econometrics where a random intercept is included in the model, one has to adjust \( A_i \) – if the random intercept is the \( i \)th column of \( W_i \), the column of \( A_i \) which is an \( i \)th unit vector should be dropped. For example, for the simple matrix \( A_i = I \otimes (1, \tilde{y}_{i0}) \), if the intercept is the first column of \( W_i \), we drop the \( \tilde{r}_{ij} \) column.\(^{1}\) \( E(\beta_i | y_{i0}, X_i^*, W_i, s, \gamma, D) \) need not necessarily be modeled as a linear function. In such cases, one can add higher order terms or other summaries of the covariates to the matrix \( A_i \).
first column of $A_i$; if the intercept is the second column of $W_i$, we drop the third column of $A_i$, and so on. Hence, regardless of whether $W_i$ includes an intercept, this specification of $A_i$ leads to a matrix $W_iA_i$ which does not include an intercept. Similarly, if $s_i$ is the $j$th column of $W_i$, for identification purposes the column of $A_i$ which is a $j$th unit vector should be dropped so that the product $W_iA_i$ does not contain $s_i$. The reason for these choices becomes apparent below, where we can write (2), after marginalizing $\bar{\gamma}_i$ using the distribution in (3), as

$$
\Pr(y_i|y_{i0}, \beta, g_i, D) = \int_{B_i} \cdots \int_{B_i} N_T(z_i|X_i\beta + g_i, V_i)dz_i,
$$

where

$$
X_i = \begin{pmatrix} X_i^* & W_iA_i & L \end{pmatrix},
$$

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

$$
V_i = I + W_iDW_i'.
$$

Under the above specification, neither $s_i$ nor an intercept are present in the matrix $X_i$, achieving the dual purpose of allowing for useful modeling of the inter-cluster heterogeneity, but at the same time resolving the identification problem under a general, unrestricted $g(\cdot)$. It should 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 $X_i^*$ or $W_i$, as long as these vary among clusters. One should be aware, however, that their simultaneous inclusion into $A_i$ to model correlation with a random intercept leaves the likelihood unidentified (since $W_iA_i$ will cause $X_i$ to contain two or more identical columns across all $i$).

The hierarchical structure of the model is completed by the introduction of (conjugate) prior densities for the model parameters $\beta$ and $D$. Gaussian priors are used to summarize the prior information about the $k$-vector $\beta$, while a Wishart prior is used for the $q \times q$ matrix $D^{-1}$:

$$
\beta \sim N_k(\beta_0, B_0), \quad D^{-1} \sim W_q(\rho_0, R_0).
$$

The specification of the prior for the function $g(\cdot)$ will be discussed in the next subsection.
From (2) and (6) it can be seen that the models here have a conditionally Gaussian structure given the unobserved effects, however, marginalization over those introduces correlation in the response variables. Then the joint density is no longer spherical, allowing the joint distribution of the elements of $y_i$ to be more general (not a product of probits), and allowing the unit-specific effects $\beta_i$ to be influenced by covariates and by the initial conditions. Furthermore, with the introduction of modeling for a heavy-tailed distribution for $\{b_i\}$ (discussed in Section 6) the model becomes more flexible and is no longer Gaussian.

2.2 The Prior on $g(\cdot)$

We place a Markov process smoothness prior on the function $g(\cdot)$. Similar priors have been discussed and used, among others, in Shiller (1973, 1984), Gersovitz and MacKinnon (1978), Besag et al. (1995), Fahrmeir and Tutz (1997, Chapter 8), Müller et al. (2001), and Fahrmeir and Lang (2001). 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. The prior we consider is also closely related to the integrated Wiener process prior in Wahba (1978) and the B-spline prior in Silverman (1985).

In particular, suppose that the observations on the covariate $s$ determine the $m \times 1$ design point vector $v$ with entries equal to the unique ordered values of $s$ with

$$v_1 < \ldots < v_m,$$

and with

$$g = (g(v_1), \ldots, g(v_m))' = (g_1, \ldots, g_m),^\perp,$$

being the corresponding function evaluations. In our implementation, the function evaluations are modeled as resulting from the realization of a $p$th order Markov process, with the specification aimed at penalizing rough functions $g(\cdot)$. Defining $h_t = v_t - v_{t-1}$, a second order prior can be
derived from the second order differences
\[
\frac{g_t - g_{t-1}}{h_t} = \frac{g_{t-1} - g_{t-2}}{h_{t-1}}, \quad (t \geq 3),
\]
leading to a second order random walk specification
\[
g_t = \left(1 + \frac{h_t}{h_{t-1}}\right) g_{t-1} + \frac{h_t}{h_{t-1}} g_{t-2} + u_t, \quad u_t \sim N \left(0, \tau^2 h_t\right),
\]
where \(\tau^2\) is a smoothness parameter. Small values of \(\tau^2\) produce smoother functions; larger values allow the function to be more flexible and interpolate the data more closely. To complete the specification of the smoothness prior, we provide a distribution for the initial states of the random walk process
\[
\begin{pmatrix} g_1 \\ g_2 \end{pmatrix} | \tau^2 \sim N \left( \begin{pmatrix} g_{10} \\ g_{20} \end{pmatrix}, \tau^2 G_0 \right),
\]
where \(G_0\) is a \(2 \times 2\) symmetric positive definite matrix. The prior on the initial conditions (10) induces a prior on linear functions of \(v\), 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).

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 \(g\) results after rewriting the Markov process in an undirected random field form. To see this, note that after defining
\[
H = \begin{pmatrix}
1 & 1 \\
-\frac{h_t}{h_{t-1}} & \left(1 + \frac{h_t}{h_{t-1}}\right) & 1 \\
& \ddots & \ddots & \ddots \\
& & -\frac{h_m}{h_{m-1}} & \left(1 + \frac{h_m}{h_{m-1}}\right) & 1
\end{pmatrix},
\]
and
\[
\Sigma = \begin{pmatrix} G_0 \\ h_3 \\
& \ddots \\
& & h_m
\end{pmatrix},
\]
10
the global smoothness representation of the second order Markov process prior equivalent to (9) and (10) becomes

\[ g | \tau^2 \sim N \left( g_0, \tau^2 K^{-1} \right), \]  

where \( g_0 = H^{-1} \tilde{g} \), with \( \tilde{g} = (g_{10}, g_{20}, 0, ..., 0)' \) (\( g_0 \) can equivalently be derived by taking recursive expectations of (9) starting with the mean in (10)), and the penalty matrix \( K \) is given by \( K = H^T \Sigma^{-1} H \). A key feature of the prior in (11) is that it is proper. This 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 \( K \) 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 \( g \) is defined conditional of the hyperparameter \( \tau^2 \), in the next level of the hierarchy we specify the prior distribution

\[ \tau^2 \sim IG \left( \frac{\nu_0}{2}, \frac{\delta_0}{2} \right). \]  

In setting the parameters \( \nu_0 \) and \( \delta_0 \) it is helpful to use the well known mapping between the mean and variance of the inverse gamma distribution and the parameters \( \nu_0 \) and \( \delta_0 \). For example, if \textit{a priori} we want \( \tau^2 \) to have mean and standard deviation of 0.5 and 0.1, respectively, we should use the prior \( \tau^2 \sim IG (54/2, 26/2) \).

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 \( K \) 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 \( \sum_i T_i \) 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 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 AND COVARIATE EFFECTS

Marginalization over the latent \( \{b_i\} \) and \( \{z_i\} \) renders the likelihood function of the model intractable analytically, as the integral in the likelihood contribution (6) is not available in closed form. However, estimation of the model is greatly facilitated by the use of the latent variable augmentation approach of Albert and Chib (1993), where the latent variables are explicitly included in the MCMC sampler, and are updated at every step. They are marginalized out (much more easily) instead by only collecting the remaining parameters from the Markov chain (as those follow the marginal distribution). Latent variable augmentation is also useful because given \( \{z_i\} \) (and \( \{b_i\} \)) the posterior updates are similar to the standard Bayes updates for continuous data models. In accordance with (6), we thus write

\[
z_i | y_{i0}, \beta, b_i, g_i, D \sim N_T(X_i \beta + g_i, V_i), \quad i = 1,..., n,
\]

(13)

where \( V_i = I_{n_i} + W_i D W_i' \) and the observed responses \( y_{it} \) are given by

\[
y_{it} = \begin{cases} 1 & \text{if } z_{it} > 0 \\ 0 & \text{if } z_{it} \leq 0 \end{cases},
\]

which induces a truncation for \( z_{it} \) at zero (truncation is at the left if \( y_{it} = 1 \), and at the right if \( y_{it} = 0 \)). Now given the \( \{z_i\} \), and defining the vectors \( b = (b'_1, ..., b'_n)' \) and \( z = (z'_1, ..., z'_n)' \), the matrix

\[
X = \begin{bmatrix} X_1 \\ \vdots \\ X_n \end{bmatrix},
\]
and the block-diagonal matrix

\[ W = \begin{bmatrix} W_1 & \cdots & W_n \end{bmatrix}, \]

one can write the model in stacked form as

\[ z = X\beta + Ng + Wb + \varepsilon, \quad \varepsilon \sim N\left(0, \sigma^2 I\right), \tag{14} \]

where \( N \) is an *incidence matrix* of dimension \( (\sum_{i=1}^n T_i) \times m \), with entries \( N_{ij} = 1 \) if \( s_i = v_j \) and 0 otherwise. In other words, the \( i \)th row of \( N \) contains a 1 in the position where the observation on \( s \) for that row matches the design point from the vector \( v \), and all remaining elements are zeros, so that \( s = Nv \).

The algorithm described below is derived from (13) and (14), and is based on an efficient blocking scheme which was proposed by Chib and Carlin (1999) for sampling the fixed and random effects in one block in order to improve the mixing of Markov chain. We note that with data augmentation, there are now two sets of latent variables in the sampler, \( \{z_{it}\} \) and \( \{b_i\} \), 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 \( \{z_i\}|y,D,\beta \) marginal of \( \{b_i\} \) by drawing for \( i \leq n, \ t \leq T_i \)

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

where \( \mu_{it} = E(z_{it}|z_{i(-t)},\beta,g_i,V_i), \ v_{it} = Var(z_{it}|z_{i(-t)},\beta,g_i,V_i), \) and \( V_i = I_{n_i} + W_iDW_i' \).

2. Sample \( \beta,\{b_i\}|y,D,\{z_{it}\},g_i \) in one block by drawing

(a) \( \beta|y,D,\{z_{it}\},g \sim N_k(\hat{\beta},B_n), \) where \( \hat{\beta} = B_n(B_0^{-1}\beta_0 + \sum_{i=1}^n X_i'V^{-1}(z_i - Ng)) \) and \( B_n = (B_0^{-1} + \sum_{i=1}^n X_i'V^{-1}X_i)^{-1} \) are the usual updates based on the complete data;
(b) \( \{b_i\} | y, D, \{z_{it}\}, \beta, g \) using \( b_i \sim N(\hat{b}_i, C_i) \), where \( \hat{b}_i = C_i W'_i (z_i - X_i \beta - g_i) \)
and \( C_i = (D^{-1} + W'_i W_i)^{-1} \).

3. Sample \( D^{-1} | \{b_i\} \sim W_p \left\{ \rho_0 + n, (R_0^{-1} + \sum_{i=1}^n b_i b_i')^{-1} \right\} \).

4. Sample \( g | y, \beta, \{b_i\}, \tau^2, \{z_{it}\} \sim N(\hat{g}, G) \), where \( G = (K/\tau^2 + N'N)^{-1} \), and where
\[
\hat{g} = G \left( K g_0/\tau^2 + N'(z - X \beta - Wb) \right).
\]
Please refer to Remark 1 below for important notes on the sampling in this step.

5. Sample \( \tau^2 | g \sim IG \left( \frac{\nu_0 + m}{2}, \frac{\delta_0 + (g - g_0)' K (g - g_0)}{2} \right) \).

**Remark 1.** In sampling \( g \), one should note that \( N'N \) is a diagonal matrix whose \( j \)th diagonal entry equals the number of values in \( s \) corresponding to the design point \( v_j \). Since \( K \) and \( N'N \) are banded, \( G^{-1} \) is banded as well. Thus sampling of \( g \) need not include an inversion to obtain \( G \)
and \( \hat{g} \). The mean \( \hat{g} \) can be found instead by solving \( G^{-1} \hat{g} = (K g_0/\tau^2 + N'(z - X \beta - Wb)) \),
which is done in \( O(m) \) operations by back substitution. Also, let \( P'P = G^{-1} \), where \( P \) is the
Cholesky decomposition of \( G^{-1} \) and is also banded. To obtain a random draw from \( N(\hat{g}_i, G_i) \)
efficiently, sample \( u \sim N(0, I) \), and solve \( P x = u \) for \( x \) by back substitution. It follows that
\( x \sim N(0, G) \). Adding the mean \( \hat{g}_i \) to \( x \), one obtains a draw \( g \sim N(\hat{g}, G) \).

We note that the MCMC approach to estimating \( \tau^2 \) in this hierarchical model offers an
alternative to 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. 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.

As with most binary data models, the parameter estimates for the model above are not
directly interpretable, due to the nonlinearity induced by the link function. Because of this
nonlinearity the effect of a given covariate \( x_j \) on the probability of response will be influenced by the remaining covariates and parameters. Moreover, due to the dynamic structure of the model, a change in \( x_j \) affects not only the contemporaneous response, but also future values of the dependent variable. Hence, to focus on the effect of \( x_j \) on contemporaneous and future \( y_{it} \), it is useful to marginalize out the covariates and parameters, thus obtaining the average covariate effect of \( x_j \). The dynamic effect on future values of \( y_{it} \) is complicated by the nondifferentiability induced by the threshold crossing nature of binary models – we thus calculate the effect via a simulation-based approach using the output from Algorithm 1. To enhance understanding, suppose the canonical model for a new individual \( i \) is given by

\[
 z_{it} = x'_{it} \beta + w'_{it} b_i + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it},
\]

where \( x'_{it} = (x'_{it}, w'_{it} \Lambda_i) \), \( \beta = (\delta', \gamma')' \), and we are interested in the effect of a particular \( x \), say \( x_1 \), on contemporaneous and future \( y_{it} \). Splitting \( x'_{it} \) and \( \beta \) accordingly, we re-write the above model as

\[
 z_{it} = x'_{1it} \beta_1 + x'_{2it} \beta_2 + w'_{it} b_i + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2} + \varepsilon_{it}.
\]

The average covariate effect can then be analyzed from a predictive perspective applied to this new individual \( i \). With this in mind, suppose that one thinks of setting \( x_{1i1} \) to the value \( x_{1j1} \). 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 \( \{x_{2it}\}, b_i \), and \( \theta = (\beta, \phi, g, D, \tau) \) given the current data \( y = (y_1, ..., y_n) \). A practical procedure is to marginalize out the covariates as a Monte Carlo average using their empirical distribution, while \( \theta \) will be integrated out with respect to the posterior distribution \( \pi(\theta | y) \). Of course \( b_i \) is independent of \( y \) and hence can be integrated out of the joint distribution of \( \{z_{i1}, ..., z_{iT_i}\} \) analytically using the distribution \( N(0, D) \), without recourse to Monte Carlo. Therefore, the
goal is to obtain a sample of draws from the distribution

\[
[z_{i1}, \ldots, z_{iT}, y_{i0}, \mathbf{x}\mathbf{1}_{it}] = \int \left[[z_{i1}, \ldots, z_{iT}, y_{i0}, \mathbf{x}\mathbf{1}_{it}], \{w_{it}\}, \{s_{it}\}; \theta] \pi(\{x_{2it}\}, \{w_{it}\}, \{s_{it}\}) \pi(\theta|y) d\{x_{it}\} d\{w_{it}\} d\{s_{it}\} d\theta.
\]

There are four possible initial conditions for this subject which means that there are four possible joint distributions for a given value of \( x_{1i1} \). Consider, for example, the case where \( y_{i0} = (0, 0)' \). A sample from the above predictive distribution can be obtained by the method of composition applied in the following way. Find all the individuals \( N_{00} = \{i : y_{i0} = (0, 0)'\} \). Randomly draw one individual \( i^* \) from this set and extract the sequence of covariate values \( \{x_{2i1}, x_{2i2}, \ldots, x_{2iT}\} \). Sample a value for \( \theta \) from the posterior density and sample \( \{z_{i1}, \ldots, z_{iT}\} \) jointly from \( [z_{i1}, \ldots, z_{iT}, y_{i0}, \mathbf{x}\mathbf{1}_{it}], \{w_{it}\}, \{s_{it}\}, \theta] \), constructing the \( y_{it} \) in the usual way. Repeat this for other draws from the posterior distribution to obtain the predictive probability mass function of \( (y_{i1}, \ldots, y_{iT}) \). Repeat this analysis for a different value of \( x_{1i1} \), say \( x_{1i1}' \), and compute the predictive mass function as above. The difference in pointwise probabilities gives the effect of \( x_1 \) as it is changed from \( x_{1i1} \) to \( x_{1i1}' \). Finally, repeat these steps for the other three possible initial conditions. This approach can similarly be applied to other elements of \( x_{it} \), as well as to elements of \( w_{it} \). In addition, the predictive horizon can be extended further into the future, but at the cost of making potentially strong assumptions about the covariates.

4 MODEL COMPARISON

A central issue in the analysis of statistical data is model formulation, since the appropriate specification is rarely known and is thus 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 or nonparametric model), or the distributional assumptions. We thus consider the general situation

\footnote{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.}
where competing hypotheses about the data $y = (y_1, \ldots, y_n)$ are captured in a collection of models $\{M_1, \ldots, M_L\}$, with each model $M_i$ characterized by a model-specific parameter vector $\theta_i \in S_i \subseteq \mathbb{R}^{k_i}$ of dimension $k_i$ and sampling density $f(y|M_i, \theta_i)$. Bayesian model selection proceeds by pairwise comparison of the models in $\{M_i\}$ through their posterior odds ratio, which for any two models $M_i$ and $M_j$ is written as

$$
\frac{\Pr(M_i|y)}{\Pr(M_j|y)} = \frac{\Pr(M_i)}{\Pr(M_j)} \times \frac{m(y|M_i)}{m(y|M_j)}
$$

where

$$
m(y|M_i) = \int f(y|M_i, \theta_i) \pi_i(\theta_i|M_i) d\theta_i
$$

is the marginal likelihood of $M_i$. The first fraction on the right hand side of (15) is known as the prior odds and the second as the Bayes factor. Because the marginal likelihood is the normalizing constant of the posterior density, one can write

$$
m(y|M_i) = \frac{f(y|M_i, \theta_i) \pi(\theta_i|M_i)}{\pi(\theta_i|y, M_i)}
$$

which is referred to as the basic marginal likelihood identity (Chib (1995)). Evaluating the right hand side of this identity at some appropriate point $\theta_i^*$ and taking logarithms one obtains the expression

$$
\log m(y|M_i) = \log f(y|M_i, \theta_i^*) + \log \pi(\theta_i^*|M_i) - \log \pi(\theta_i^*|y, M_i)
$$

from which the marginal likelihood can be estimated by finding an estimate of the posterior ordinate $\pi(\theta_i^*|y, M_i)$. Thus, the calculation of the marginal likelihood is reduced to finding an estimate of the posterior density at a single point $\theta_i^*$. For estimation efficiency the latter point is generally taken to be a high density point in the support of the posterior.

Chib (1995) provides a method to estimate the posterior ordinate $\pi(\theta_i^*|y, M_i)$ in the context of Gibbs MCMC sampling. Suppose that the parameter space is split into $B$ conveniently specified blocks, so that $\theta^* = (\theta_1^*, \ldots, \theta_B^*)$, suppressing the model index for notational convenience.
Then, by the law of total probability we have

$$\pi(\theta^* | y) = \pi(\theta^*_1 | y) \pi(\theta^*_2 | y, \theta^*_1) \cdots \pi(\theta^*_B | y, \theta^*_1, ..., \theta^*_{B-1})$$

(19)

where $\pi(\theta^*_1 | y)$ is the marginal density ordinate of $\theta_1$ and $\pi(\theta^*_B | y, \theta^*_1, ..., \theta^*_{B-1})$ is the full conditional density ordinate and the remaining ordinates are reduced conditional ordinates. Since each full conditional density is known, then the marginal density ordinate is estimated by the Rao-Blackwell device (Tanner and Wong 1987; Gelfand and Smith 1990). Next, the first reduced conditional ordinate is found by averaging the full conditional density of $\theta_2$, $\bar{\pi}(\theta_2^* | y, \theta_1^*) = M^{-1} \sum_{j=1}^{M} \pi(\theta_2^* | y, \theta_1^*, \theta_3^{(j)}, ..., \theta_B^{(j)})$ where $\{\theta_3^{(j)}, ..., \theta_B^{(j)}\} \sim \pi(\theta_3, ..., \theta_B | y, \theta_1^*)$ are $M$ draws that are obtained from a reduced MCMC run in which $\theta_1$ is fixed at $\theta_1^*$ and sampling is over $\{\theta_2, ..., \theta_B\}$, a procedure that requires no new programming. Subsequent reduced ordinates are estimated in the same way by fixing additional blocks. The computational cost of this procedure is generally small when blocking is effectively done and a few reduced runs are required, as is possible in many practical problems. We note that while it is true that the identity (17) can also be written as $m(y | M_l) = f(y | M_l, \theta_l, z_l) \pi(\theta_l, z_l | M_l) / \pi(\theta_l, z_l | y, M_l)$ when latent variables $z_l$ are present, that form of the identity is not very useful, because the dimension of $z_l$ may be very large (easily running into the hundreds or thousands). We will therefore generally integrate out any such parameters before applying (17). 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 $y$ in (15)).

The choice of suitable decomposition in (19) determines an appropriate balance between computational and statistical efficiency. To see this, consider the case when a large dimensional block is placed towards the front of the decomposition in (19). Because this block is held fixed in subsequent reduced runs, the computational demands are lower. This, however, may increase the variability in the Rao-Blackwellization step, where the full-conditional density average for the large block is taken over a conditioning set which changes with every iteration. Alternatively,
if the large dimensional block is placed towards the end in (19), the Rao-Blackwell average will be more stable as now more blocks in the conditioning set stay fixed; this strategy leads to higher statistical efficiency but comes at a higher computational cost, since a large dimensional block (rather than a different block of lower dimension) is simulated in all of the preceding reduced runs. Because of these concerns, for the estimation of the marginal likelihood of the model we presented in Section 2, we use the decomposition

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

marginalized over the latent variables \(\{z_i\}\) and \(\{b_i\}\). Since \(g\) may potentially be of dimension up to the total number of observations in the sample, it is placed last in (20). We note, however, that because the simulation algorithm for \(g\) is \(O(m)\) as discussed in Section 3, this particular choice comes at a small increase in computational cost, while the statistical efficiency benefits may be substantial, especially for large-dimensional \(g\). We compared several alternative decompositions, some of which required fewer reduced runs (and were thus slightly faster), but in line with the arguments above, we found that the decomposition in (20) reduced variability in the marginal likelihood estimate, and worked well for both small- and large-dimensional \(g\). Further analysis of the model comparison method is taken up in Chib and Jeliazkov (2002), where marginal likelihoods are estimated for binary and continuous data models, and the correctness of the estimates is verified for cases in which answers are available by alternative estimation methods.

The implementation of (18), marginal of \(\{z_i\}\) and \(\{b_i\}\), requires the likelihood ordinate \(f(y|D^*, \beta^*, \tau^{2*}, g^*)\). To obtain this ordinate, we use the Geweke, Hajivassiliou, and Keane (GHK) method, which provides estimates of the likelihood contributions (6) at the values \((D^*, \beta^*, \tau^{2*}, g^*)\). For a detailed discussion of this method and its properties, see Börsch-Supan and Hajivassiliou (1993) and Keane (1994). The method is based on writing \(V_i = LL'\), where \(L\) is a lower triangular Cholesky factorization, and making a change of variable from \(z_i\) to \(\varepsilon_i\)
where \( z_i = X_i \beta + g_i + L \varepsilon_i \). Then

\[
\Pr(y_i | y_{i0}, \beta^*, g^*_i, D^*) = \int_{B_{Ti}} \cdots \int_{B_{i1}} N_T(z_i | X_i \beta^* + g^*_i, V_i) dz_i
\]

\[
= \int_{c_{Ti}}^{d_{Ti}} \cdots \int_{c_{i1}}^{d_{i1}} N_T(u|0, I) du,
\]

where

\[
c^*_{it} = \frac{c_{it} - x_{it} \beta^* - g_{it}^* - \sum_{k=1}^{l} t_{ik} \varepsilon_{ik}}{l_{it}}, \quad d^*_{it} = \frac{d_{it} - x_{it} \beta^* - g_{it}^* - \sum_{k=1}^{l} t_{ik} \varepsilon_{ik}}{l_{it}},
\]

and \( c_{it} \) and \( d_{it} \) denote the lower and 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.

5 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 \( E \{ g(v) | y \} \) 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 \( \{T_i\} \).
Data is simulated from the model in (1) and (3), using 1, 2, and 3 lags, a single fixed effect covariate $X$, and 1 or 2 individual effect covariates $W$ (including a random intercept). $X$ and $W$ contain independent standard normal random variables, and we use $\delta = 1$, $\gamma = 1$, $\phi = 0.5*1$, and $D = 0.2*I$. 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

\[
MSE = \frac{1}{m} \sum_{j=1}^{m} \{\hat{g}(v_j) - g(v_j)\}^2.
\]  

(21)

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. 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 \(\tau^2\), 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 \(\phi\) 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 \(q = 1\).

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 ($\tau^2$ and $D$) 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 $\tau^2$ 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 $D$ determine a balance between intra- and inter-cluster variation. While in large samples the effects of the assumed priors on the parameter

\begin{table}[h]
\centering
\begin{tabular}{cccccc}
\hline
Clusters & Lags & Random effects & $g_1$ & $g_2$ & $g_3$ \\
\hline
\multirow{2}{*}{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) \\
\hline
\multirow{2}{*}{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) \\
\hline
\multirow{2}{*}{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) \\
\hline
\end{tabular}
\caption{Average mean squared errors based on 10 samples, with estimated standard errors in parentheses.}
\end{table}
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 $\tau^2$. In one case the prior on $\tau^2$ is such that $E(\tau^2) = 0.5$ and $SD(\tau^2) = 0.1$; in the second case $E(\tau^2) = 0.001$ and $SD(\tau^2) = 0.001$. 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.

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 $\tau^2$ appear to be estimated well and the sample is characterized by low autocorrelations. While it can be seen that $D$ 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 $D$ more accurately. The slower mixing occurs because $D$ is a
Figure 3: Effect of $\tau^2$ on the estimates of $g$ (solid lines) for two sample sizes: $n = 250$ in the first column, and $n = 1000$ in the second. Since $\tau^2$ is large in the first row, the estimated function is less smooth; a small $\tau^2$ in the second row leads to oversmoothing. The confidence bands (dashed lines) are tighter in the second column because of the larger sample size.

Parameter at the second level of the modeling hierarchy and depends on the data only indirectly through $\{b_i\}$ (i.e. given $\{b_i\}$, $D$ does not depend on $\{z_i\}$ and $y$). Since $\{b_i\}$ are not well identified in smaller clusters, when only a few observations are available to identify the cluster-specific effects, and because learning about $D$ occurs from the inter-cluster variation of $\{b_i\}$, $D$ also suffers from weak identification when cluster sizes are small. 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 $\{b_i\}$ better, allowing for inter-cluster variation to be captured more precisely. In line with the arguments above, Table 2 shows that the inefficiency factor for $D$ drops considerably (the other inefficiency factors stay within a similar range). The improvement in the sampling
Figure 4: Posterior samples and autocorrelations for the parameters of a semiparametric model with one fixed effect, one random effect, and one lag ($T_i = 7, i = 1, ..., n$).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$T_i = 7$</th>
<th>$T_i = 12$</th>
<th>$T_i = 17$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>13.683</td>
<td>10.183</td>
<td>12.553</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>9.617</td>
<td>6.155</td>
<td>7.247</td>
</tr>
<tr>
<td>$\phi$</td>
<td>12.111</td>
<td>8.310</td>
<td>9.031</td>
</tr>
<tr>
<td>$D$</td>
<td>24.002</td>
<td>13.276</td>
<td>9.413</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td>10.654</td>
<td>5.260</td>
<td>8.740</td>
</tr>
</tbody>
</table>

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 $n = 500$.

of $D$ 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 $T_i = 17$.

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 $D$ are sampled with somewhat higher inefficiency factors. Here again, however, Table 3 shows that as the cluster sizes increase, resulting in better identification of $\{b_i\}$, the inefficiency factors for the elements of the heterogeneity matrix $D$ drop noticeably.
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 $D$ benefits strongly from the availability of larger cluster sizes.

6 MODEL EXTENSIONS AND RELATED FITTING TECHNIQUES

We mentioned in Section 2 that it is possible to introduce a heavy-tailed distribution for $b_i$, allowing for a more flexible non-Gaussian model. In particular, to allow for a Student-$t$ distribution of the random effects $b_i$, we add the following hierarchical level to the model (replacing the distributional assumption in (4)), for $i = 1, ..., n$,

$$b_i | \lambda_i \sim N \left( 0, \lambda_i^{-1} D \right), \quad \lambda_i \sim G \left( \nu/2, \nu/2 \right).$$

(22)

It can then be shown that marginalized over $\{ \lambda_i \}$, the $\{ b_i \}$ have a multivariate $t$ distribution. Typically $\nu$ will be fixed at the outset, with this specification allowing for fat-tailed distributions.
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.

for low values of $\nu$, and an approximately normal distribution if $\nu$ is large. An algorithm for the sampling of a model with multivariate $t$ random effects is presented next. We note that the model now contains three sets of latent variables $\{z_{it}\}, \{b_i\},$ and $\{\lambda_i\}$ – and that conditional on $\{\lambda_i\}$ the sampler has a structure similar to the one in Algorithm 1.

Algorithm 2 State Dependence Model with Multivariate $t$ Random Effects: MCMC Implementation

1. Sample $\{z_{it}\}|y,D,\beta,\{\lambda_i\}$ marginal of $\{b_i\}$ by drawing for $i \leq n$, $t \leq T_i$

   
   $z_{it} \sim \begin{cases} 
   \mathcal{T}_N(0,\infty)(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 1 \\
   \mathcal{T}_N(-\infty,0)(\mu_{it}, v_{it}) & \text{if} \quad y_{it} = 0 
   \end{cases}
   
   $\mu_{it} = E(z_{it}|z_{i(-t)},\beta, V_i)$, $v_{it} = \text{Var}(z_{it}|z_{i(-t)},\beta, V_i)$, and $V_i = I_{n_i} + W_i\lambda_i^{-1}DW_i$.

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

   (a) $\beta|y,D,\{z_{it}\},\{\lambda_i\} \sim \mathcal{N}_k(\hat{\beta}, B_n)$ where $\hat{\beta} = B_n(B_0^{-1}\beta_0 + \sum_{i=1}^n X_i'V_i^{-1}(z_i - Ng))$

   and $B_n = (B_0^{-1} + \sum_{i=1}^n X_i'V_i^{-1}X_i)^{-1}$;

   (b) $\{b_i\}|y,D,\{z_{it}\},\beta,\{\lambda_i\},g$ using $b_i \sim \mathcal{N}_q(\hat{b}_i, C_i)$, where $\hat{b}_i = C_iW_i'(z_i - X_i\beta - g_i)$

   and $C_i = (\lambda_iD^{-1} + W_i'W_i)^{-1}$.
3. Sample \( \{ \lambda_i \} \mid \{ b_i \} \) from \( \lambda_i \sim G((\nu + q)/2, (\nu + b_i' D^{-1} b_i)/2) \).

4. Sample \( g \mid y, \beta, \{ b_i \}, \tau^2, \{ z_i \} \sim N(\hat{g}, G) \), where \( G = (K/\tau^2 + N'N)^{-1} \), and where \( \hat{g} = G(Kg_0/\tau^2 + N'(z - X\beta - Wb)) \).

5. Sample \( D^{-1} \mid \{ b_i \}, \{ \lambda_i \} \sim W_p \left\{ \rho_0 + n, (R_0^{-1} + \sum_{i=1}^n \lambda_i b_i b_i')^{-1} \right\} \).

6. Sample \( \tau^2 \mid g \sim IG \left( \frac{m+n}{2}, \frac{\delta_0 + (g - g_0)^2}{2} \right) \).

It is also possible to place a prior on \( \nu \), so that it becomes a parameter to be estimated. For the application of such models it necessary to have a large number of observations in the tails of the distribution, so as to identify \( \nu \) reasonably well. For further details on how this may be approached, see Albert and Chib (1993).

The marginal likelihood of this model is estimated using the framework described in Section 4, and is based on decomposition (20). The difference here is that the full-data likelihood will have to be marginalized over \( \{ \lambda_i \}, \{ b_i \}, \) and \( \{ z_i \} \), to obtain the likelihood ordinate \( f \left( y \mid D^*, \beta^*, \tau^{2*}, g^* \right) \). This ordinate can be found by importance estimation (Geweke (1989)), evaluating the average

\[
\frac{1}{M} \sum_{j=1}^{M} \frac{f \left( y \mid D^*, \beta^*, \tau^{2*}, g^*, \{ b_i \}^{(j)}, \{ \lambda_i \}^{(j)} \right) \prod_{i=1}^{n} \left\{ N_{q} \left( b_i^{(j)} | 0, D^* / \lambda_i^{(j)} \right) \right\} \frac{G \left( \lambda_i^{(j)} | \nu/2, \nu/2 \right)}{h \left( \{ b_i \}^{(j)} | y, D^*, \beta^*, \tau^{2*}, g^* \right) \prod_{i=1}^{n} G \left( \lambda_i^{(j)} | \nu/2, \nu/2 \right)} \}
\]  
(23)

with draws \( \lambda_i \sim G(\nu_1/2, \nu_2/2) \), for \( i = 1, \ldots, n \), and \( \{ b_i \} \sim h \left( \{ b_i \} | y, D^*, \beta^*, \tau^{2*}, g^* \right) \). The importance density \( h(\cdot) \) is specified as a the product of \( n \) multivariate-\( t \) densities with 10 degrees of freedom, and we use the draws of \( \{ b_i \} \) and \( \{ \lambda_i \} \) from the last reduced run to obtain the necessary means and variances, as well as \( \{ \nu_1 \} \) and \( \{ \nu_2 \} \). Given the draws for \( \{ b_i \} \) and \( \{ \lambda_i \} \) from the importance density, the integral over \( \{ z_i \} \) to obtain \( f \left( y \mid D^*, \beta^*, \tau^{2*}, g^*, \{ b_i \}^{(j)}, \{ \lambda_i \}^{(j)} \right) \) in the numerator of (23) is available analytically. Importance estimation can also be used to find the likelihood ordinates of models with mixture-of-normals, logit, or multivariate-\( t \) link functions,
by similarly constructing importance densities for the additional latent variables that are used in the sampling of such models (for details on sampling the additional latent variables, see Albert and Chib (1993) and Wood and Kohn (1998)).

While this paper is concerned with the modeling and estimation of a flexible semiparametric model with state dependence, we note that the estimation of models with serial correlation (where the errors are assumed to follow an autoregressive process) can be done by using the algorithms in Chib and Greenberg (1998) who analyze multivariate probit models for correlated binary data. The correlation structures considered by Chib and Greenberg (1998) allow for general as well as for structured correlation matrices, and hence can be applied to the Toeplitz structures which emerge in the analysis of models with serial correlation. Details on the MCMC analysis of probit models for correlated binary data are given in Chib and Greenberg (1998), and related model selection issues for that model are discussed in Chib and Jeliazkov (2001).

In a final remark we note that while we model the lag coefficients as common across the clusters, a unit-specific representation for the lag coefficients is also possible with slight changes in notation, but without any changes to the algorithm.

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 4. 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.\(^3\)

Similar data have been analyzed by Chib and Greenberg (1998), who estimated multivariate

\(^3\)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.
Table 4: Variables in the women’s labor force participation application. The dependent variable is the woman’s labor force participation 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).

Table 4: Variables in the women’s labor force participation application. The dependent variable is the woman’s labor force participation 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).

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

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 4 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 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 \textit{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 be a dangerous practice, since 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”.

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 ($M_1$) is given by:

$$\Pr(y_{it} = 1|\theta, \beta_1) = F\left(x_{it}'\delta + w_{it}'\beta_1 + g(s_{it}) + \phi_1 y_{i,t-1} + \phi_2 y_{i,t-2}\right),$$

$$\beta_1 = A_i\gamma + b_i, \quad b_i \sim N_3(0, D),$$

where $y_{it} = WORK_{it}, \theta = (\delta, g, \phi_1, \phi_2, \gamma, D), x_{it}' = (RACE, EDU_{it}, \ln(INC_{it})), s_{it} = AGE_{it}, w_{it}' = (1, CH2_{it}, CH5_{it}),$ and

$$A_i = \begin{pmatrix} \bar{y}_{i0} & 1 & \ln(INC_{i}) \\ 1 & \bar{y}_{i0} & \ln(INC_{i}) \end{pmatrix}. $$

Because of its features, model $M_1$ 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.

The parameter estimates for $M_1$ are presented in Table 5. Interpretation of the estimates in Table 5 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

---

4 In our sample, $\text{Corr}(EDU, INC) = 0.30$ and $\text{Corr}(EDU, \ln(INC)) = 0.32$. 

---
<table>
<thead>
<tr>
<th>Parameter</th>
<th>Covariate</th>
<th>Mean</th>
<th>SD</th>
<th>Median</th>
<th>Lower</th>
<th>Upper</th>
<th>Ineff</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\delta$</td>
<td>$RACE$</td>
<td>0.170</td>
<td>0.080</td>
<td>0.169</td>
<td>0.014</td>
<td>0.329</td>
<td>7.012</td>
</tr>
<tr>
<td></td>
<td>$EDU$</td>
<td>0.087</td>
<td>0.015</td>
<td>0.086</td>
<td>0.057</td>
<td>0.117</td>
<td>23.189</td>
</tr>
<tr>
<td></td>
<td>$\ln (INC)$</td>
<td>-0.190</td>
<td>0.048</td>
<td>-0.189</td>
<td>-0.286</td>
<td>-0.098</td>
<td>16.484</td>
</tr>
<tr>
<td>$\gamma$</td>
<td>$\bar{y}_{i0}$</td>
<td>1.371</td>
<td>0.173</td>
<td>1.365</td>
<td>1.047</td>
<td>1.724</td>
<td>27.802</td>
</tr>
<tr>
<td></td>
<td>$CH_2$</td>
<td>0.142</td>
<td>0.312</td>
<td>0.144</td>
<td>-0.479</td>
<td>0.747</td>
<td>5.414</td>
</tr>
<tr>
<td></td>
<td>$(CH_2)(\bar{y}_{i0})$</td>
<td>-0.245</td>
<td>0.161</td>
<td>-0.248</td>
<td>-0.556</td>
<td>0.077</td>
<td>19.356</td>
</tr>
<tr>
<td></td>
<td>$(CH_2)(\ln (INC_{i1}))$</td>
<td>-0.135</td>
<td>0.093</td>
<td>-0.135</td>
<td>-0.318</td>
<td>0.046</td>
<td>6.230</td>
</tr>
<tr>
<td></td>
<td>$CH_5$</td>
<td>0.868</td>
<td>0.273</td>
<td>0.867</td>
<td>0.339</td>
<td>1.416</td>
<td>8.358</td>
</tr>
<tr>
<td></td>
<td>$(CH_5)(\bar{y}_{i0})$</td>
<td>-0.351</td>
<td>0.127</td>
<td>-0.350</td>
<td>-0.606</td>
<td>-0.103</td>
<td>14.530</td>
</tr>
<tr>
<td></td>
<td>$(CH_5)(\ln (INC_{i1}))$</td>
<td>-0.221</td>
<td>0.081</td>
<td>-0.221</td>
<td>-0.380</td>
<td>-0.063</td>
<td>8.139</td>
</tr>
<tr>
<td>$\phi$</td>
<td>$y_{i,t-1}$</td>
<td>1.213</td>
<td>0.071</td>
<td>1.213</td>
<td>1.072</td>
<td>1.348</td>
<td>15.863</td>
</tr>
<tr>
<td></td>
<td>$y_{i,t-2}$</td>
<td>0.445</td>
<td>0.071</td>
<td>0.445</td>
<td>0.308</td>
<td>0.581</td>
<td>11.470</td>
</tr>
<tr>
<td></td>
<td>$vech (D)$</td>
<td>0.540</td>
<td>0.129</td>
<td>0.528</td>
<td>0.319</td>
<td>0.828</td>
<td>38.481</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.043</td>
<td>0.096</td>
<td>-0.043</td>
<td>-0.243</td>
<td>0.133</td>
<td>45.999</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.137</td>
<td>0.071</td>
<td>0.119</td>
<td>0.046</td>
<td>0.319</td>
<td>45.617</td>
</tr>
<tr>
<td></td>
<td></td>
<td>-0.151</td>
<td>0.085</td>
<td>-0.138</td>
<td>-0.347</td>
<td>-0.019</td>
<td>43.454</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.017</td>
<td>0.049</td>
<td>0.011</td>
<td>-0.066</td>
<td>0.136</td>
<td>45.551</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.158</td>
<td>0.086</td>
<td>0.135</td>
<td>0.047</td>
<td>0.366</td>
<td>46.355</td>
</tr>
<tr>
<td>$\tau^2$</td>
<td></td>
<td>0.017</td>
<td>0.006</td>
<td>0.016</td>
<td>0.009</td>
<td>0.030</td>
<td>5.473</td>
</tr>
</tbody>
</table>

Table 5: Parameter estimates for model $M_1$. The table also reports 95% confidence intervals and inefficiency factors from 15000 MCMC iterations.

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 $N_{mn} = \{i : y_{i0} = (m,n)^t\}$, as well as overall results for the sample.\(^5\)

From Table 5 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

---

\(^5\)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: $#\{N_{00}\} = 337$, $#\{N_{01}\} = 98$, $#\{N_{10}\} = 127$, $#\{N_{11}\} = 983.$
necessary to produce economically significant changes in participation). Quite importantly, the results in Table 5 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.6

6There are many differences between the two models and the estimation techniques, and one such difference is that in M1, husband’s earnings, rather than wife’s education, are 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,
There is strong correlation between the initial conditions and the random effects in this model. In particular, the strong positive correlation between the random intercept and the initial conditions is in agreement with a human capital theory, and also with the fact that the initial observations are indicative of a woman’s tastes. The negative coefficients on \( y_{i0} \) for the effects of CH2 and CH5 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)), thus agreeing with the sign of the coefficient estimates reported above.

The estimate of the nonparametric function \( g(AGE) \) 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

The marginal likelihood for that model was lower than that of \( M_1 \) as discussed at the end of this section.
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.\footnote{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 \( g(\text{AGE}) \).}

The marginal likelihood of model \( M_1 \) 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 6. 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 \( \ln(\text{INC}) \) 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 \( M_2 \) and \( M_3 \) in Table 6.

Turning attention to the heterogeneity in the individual specific effects, we see from Table 6 that model \( M_4 \), where the conditional means of the child status effects are allowed to be correlated with \( \overline{\text{EDU}}_i \) instead of with \( \overline{\ln(\text{INC})}_i \), does not perform as well as \( M_1 \) (using both

\[ E_{i} \]
Table 6: Some alternative models in the women’s labor force participation application. In this table, we use $x_{it}^* = (RACE_{it}, EDU_{it}, \ln(INC_{it}))'$ and $w_{it} = (1, CH2_{it}, CH5_{it})'$ and except for the parametric models, the effect of age is modeled nonparametrically. Only the non-zero elements of $A_i$ are presented, with commas separating the columns in a given row, and semi-colons separating rows. The marginal likelihoods are based on MCMC runs of length 15000.

$EDU_i$ and $\ln(INC_i)$ in $A_i$ 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 $M_5$ and $M_6$. Model $M_5$ allows the individual effects to be correlated with the initial conditions but not with any covariates, while $M_6$ allows all random effects, including the intercept, to depend on $\ln(INC_i)$ and the initial conditions. Although one of the specifications is more parsimonious, while the other is less parsimonious than $M_1$, both have lower marginal likelihoods than $M_1$, illustrating that Bayes factors penalize overparameterization. Most impor-
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.

Tantamount, however, a “traditional” specification with a single unobserved effect (a random intercept) did worse than $M_1$ by a large margin as can be seen from $M_7$.

The index function of model $M_8$ is linear in age, but the model is otherwise similar to $M_1$ with 2 lags and 3 random effects (observe that $A_i$ 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 $M_9$, which uses a quadratic in age. The estimates from the two parametric models are plotted against the nonparametric estimate from $M_1$ 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 ($M_{10}$) of our baseline model had a marginal likelihood which was lower than that of $M_1$ by about 47 on the natural log scale. 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) heterogeneity in the effect of multiple covariates, and (iv) correlation between
the random effects and the covariates. The methodology developed in Sections 2-4 provides
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

In this paper we have provided a hierarchical Bayes framework for fitting semiparametric binary
panel models with state dependence. The technique relies on latent variable augmentation
(Albert and Chib (1993)) and on placing a Markov process prior on the unknown function (Shiller
(1984), Fahrmeir and Lang (2001)). 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. Key
advantages of the approach that distinguish it from previous work are the proper smoothness
prior on $g$ (allowing for finite sample model inferences), the ability to accommodate multiple
lags and multiple unobserved effects, and the ability to allow for flexible model extensions. Such extensions include models with multivariate-$t$ random effects and models with Markov process priors of different orders.

We have applied tuned MCMC methods for simulation of the posterior distribution, for estimation of the marginal likelihood, and for describing the average effect of the covariates. The fitting algorithm has a relatively low computational cost, allowing for the Bayesian analysis of large data panels. A small 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 method are practical, easily applicable, and that they 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). 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


