ELSEVIER

Journal of Econometrics 110 (2002) 67–89

JOURNAL OF
Econometrics

www.elsevier.com/locate/econbase

Semiparametric Bayes analysis of longitudinal data treatment models

Siddhartha Chib*, Barton H. Hamilton

John M. Olin School of Business, Washington University, Campus Box 1133, 1 Brookings Dr., St. Louis, MO 63130, USA

Abstract

This paper is concerned with the problem of determining the effect of a binary treatment variable on a continuous outcome given longitudinal observational data and non-randomly assigned treatments. A general semiparametric Bayesian model (based on Dirichlet process mixing) is developed which contains potential outcomes and subject level outcome-specific random effects. The model is subjected to a fully Bayesian analysis based on Markov chain Monte Carlo simulation methods. The methods are used to compute the posterior distribution of the parameters and potential outcomes. The sampled posterior output from the simulation is also used to construct treatment effect distributions at the unit level (and at other levels of aggregation), marginalized over all unknowns of the model, including the unknown distribution of responses and treatments, and treatment effects matched by treatment probability. A real data example, dealing with the wage premium associated with union membership, is considered in detail where quantities such as the average treatment effect, the treatment effect on the treated, and matched treatment effects are derived and illustrated. © 2002 Published by Elsevier Science B.V.

JEL classification: C1; C4

Keywords: Average treatment effect; Bayesian matching; Causal inference; Dirichlet process prior; Panel data

1. Introduction

This paper is concerned with the analysis of a class of models for longitudinal observational data where the main question is about the *causal effect* of a discrete treatment variable on a response given that the treatment variable is non-randomly assigned and, for any given subject, at any given time point, the outcome is observed for

* Corresponding author. Fax: +1-314-935-6359.
E-mail address: chib@olin.wustl.edu (S. Chib).

PII: S0304-4076(02)00122-7

only one level of the treatment. Examples of treatment questions include the impact of smoking by pregnant women on birth weights (Permutt and Hebel, 1989), compliance effects in drug trials (Efron and Feldman, 1991), the impact of union status on wages (Lee, 1978), and the effect of training on wages (Heckman and Robb, 1985).

One way to compute the treatment effect is based on the idea of *matching* where subjects, or groups of subjects, are matched on the basis of observed covariates, or on the basis of the so-called propensity score, with the treatment effect then calculated as the difference between the observed outcomes of the matched individuals corresponding to different levels of the treatment (Rosenbaum and Rubin, 1983; Heckman et al., 1998; Hirano et al., 2000). This method is useful provided it is the case that, conditioned on the covariates, the unobservables that influence the treatment are independent of the outcomes. If the preceding assumption does not hold, then an alternative way to proceed, as in this paper, is to specify a joint model of the treatment and the potential outcomes (i.e., the outcomes, or counter-factuals, that would have been observed for levels of the treatment not received as in Rubin (1974, 1978)) and then to compare the observed outcome with the potential outcomes for each subject. An early example of this approach is often summarized in what is called the Roy model (Roy, 1951). We note at the outset that although the treatment effect in the potential outcomes model can be identified by distributional assumptions, identification is achieved more convincingly if, additionally, there exists an exogenous covariate, an *instrument*, that affects the outcomes only through its effect on the treatment (Heckman and Robb, 1985).

Vella and Verbeek (1998) have recently considered a version of the Roy model for longitudinal data but under the assumption that the (latent) treatment and potential outcomes are conditionally Gaussian. Their model can be viewed as a longitudinal extension of the model in Bjorklund and Moffitt (1987). The model is fit by maximum likelihood. Jakubson (1991), Ridder (1992), Wooldridge (1995), Kyriazidou (1997) and Vella and Verbeek (1999) also consider longitudinal treatment models but in the sample selection context, not in the potential outcomes framework. More recently, Chib and Hamilton (2000) provide a Bayesian analysis of cross-section and longitudinal data under the assumption that the outcome and treatment distributions belong either to the multivariate-t family or the family of finite mixture of multivariate normal distributions.

The first main purpose of this paper is to further robustify a potential outcomes model for longitudinal treatment data by modeling the treatments and outcomes in a semiparametric fashion. Such an extension has not been considered before in either the Bayesian or frequentist literatures and is potentially quite useful because of the well known sensitivity of conclusions in treatment models to distributional assumptions. In the proposed model, subjects can be in different treatment states across time and intra-cluster correlation in subject-specific treatments and outcomes is captured by including treatment- and outcome-specific individual random effects that are correlated with covariates. Furthermore, the joint distribution of the treatments and potential outcomes is modeled by a semiparametric mixture of Dirichlet process components (Fergusson, 1973; Antoniak, 1974). Tiwari et al. (1988) provide one of the earliest uses of the Dirichlet process approach in econometrics. The model is then estimated by a tuned Markov chain Monte Carlo (MCMC) simulation method (Chib, 2001; Chib and Greenberg, 1995, 1996; Tiernay, 1994) based on the algorithm reported in Chib and Hamilton (2000).

The second main purpose of this paper is to develop a procedure for calculating the treatment effect by a method that we call Bayesian matching that is similar to, but distinct, from the propensity score matching scheme described above. Our approach provides a useful way of extracting and summarizing treatment effect heterogeneity that may be present in the data.

The rest of the paper is organized as follows. In Section 2 we describe the modeling framework and some implications of our setup. Section 3 contains an outline of the MCMC estimation procedure and methods for calculating treatment effects from the posterior simulation output. Sections 4 presents a real data example where quantities such as the average treatment effect, the treatment effect on the treated, and matched treatment effects are derived and illustrated. Section 5 concludes. Details of the fitting method are provided in Appendix A.

2. Modeling framework

2.1. The semiparametric panel potential outcomes model

To describe the semiparametric panel model for longitudinal continuous responses with potential outcomes, let s_{it} denote the time-dependent treatment variable taking two levels $\{0,1\}$ on the ith subject at the tth time period. Suppose that one observes n subjects in the sample, each over T_i unbalanced time periods (the process governing attrition is assumed to be ignorable), giving rise to $m=\sum_{i=1}^{n} T_i$ observations, with outcomes $\mathbf{z}_{it}^{*}=\{z_{it0},z_{it1}\}$ for the two levels of the treatment (of which one is observed). Thus, in this model, there is one variable z_{it0} representing the outcome when the treatment is not received and a different variable z_{it1} representing the outcome when the treatment is received. For the actual treatment received (say $s_{it}=l$), suppose that $y_{it}=z_{itl}$ denotes the *observed* response. Let $\mathbf{x}_{itj}:k_j\times 1$ denote the k_j dimensional covariate vector that influences the distribution of z_{itj} and let $\mathbf{w}_{it}:k_0\times 1$ denote the covariate vector that affects the distribution of the treatments. Assume that the covariate vector $\mathbf{w}_{it}$ contains *at least* one discrete or continuous covariate r_{it} that is randomly assigned (an instrument) and which is not present in $\mathbf{x}_{itj}$. If the observed and unobserved data on the ith subject is denoted by $(\mathbf{s}_i,\mathbf{z}_i^{*})$ where $\mathbf{s}_i=(s_{i1},\ldots,s_{iT_i})$ and $\mathbf{z}_i^{*}=(\mathbf{z}_{i1}^{*},\ldots,\mathbf{z}_{iT_i}^{*})$, then assume that the joint distribution of $(\mathbf{z}_i^{*},\mathbf{s}_i)$ conditioned on the covariates can be factored as

$$f(\mathbf{z}_i^{*},\mathbf{s}_i|\mathbf{W}_i,\mathbf{X}_i)=f(\mathbf{z}_i^{*}|\mathbf{X}_i)f(\mathbf{s}_i|\mathbf{z}_i^{*},\mathbf{W}_i,\mathbf{X}_i)$$

or in other words that the instrument r_{it} does not affect the marginal distribution of $\mathbf{z}_i^{*}$. This assumption is necessary in order to avoid confounding between the effect of $\mathbf{s}_i$ on $\mathbf{z}_i^{*}$ and that of the unobservables on $\mathbf{z}_i^{*}$.

The presence of potential outcomes and treatments naturally leads to a model with latent data. Let the observed treatment s_{it} be a function of a continuous-valued random variable s_{it}^* as

$$s_{it} = \begin{cases} 0 \text{ if } s_{it}^* \leqslant 0, \\ 1 \text{ if } s_{it}^* > 0 \end{cases}$$

and let the observed data y_{it} be given by

$$y_{it} = \begin{cases} z_{it0} \text{ if } s_{it} = 0, \\ z_{it1} \text{ if } s_{it} = 1. \end{cases}$$

In the longitudinal context, treatments and outcomes over time on the same subject are likely to be correlated. To model this correlation, we introduce treatment- and outcome-specific individual random effects $\mathbf{b}_i = (a_i, b_{i0}, b_{i1})$ that can be correlated with a set of covariates $\mathbf{V}_i : 3 \times d$. In particular, it is assumed that

$$\mathbf{b}_i \sim \mathrm{N}_3(\mathbf{V}_i\boldsymbol{\gamma}, \mathbf{D}),$$

where $\mathbf{D}$ is a full positive definite matrix.

Next, to allow for the possibility that the unobservables affecting the treatment are also correlated with the outcomes, even after conditioning on the covariates, we model the joint distribution of s_{it}^* and (z_{it0}, z_{it1}) in a general semiparametric fashion. Conditional on the random effects $\mathbf{b}_i$, parameters and a positive scale random variable λ_{it}, let

$$\begin{pmatrix} s_{it}^* \\ z_{it0} \\ z_{it1} \end{pmatrix}_{|\mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}} \sim \mathrm{N}_3\left(\begin{pmatrix} \mathbf{w}_{it}'\boldsymbol{\gamma} + a_i \\ \mathbf{x}_{it0}'\boldsymbol{\beta}_0 + b_{i0} \\ \mathbf{x}_{it1}'\boldsymbol{\beta}_1 + b_{i1} \end{pmatrix}, \lambda_{it}^{-1} \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \right), \tag{1}$$

or compactly as

$$\mathbf{z}_{it} | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim^{ind} \mathrm{N}_3(\mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{b}_i, \lambda_{it}^{-1}\boldsymbol{\Sigma}),$$

where $\mathbf{z}_{it} = (s_{it}^*, z_{it0}, z_{it1}) : 3 \times 1$, $\mathbf{X}_{it} = \mathrm{diag}(\mathbf{w}_{it}', \mathbf{x}_{it0}', \mathbf{x}_{it1}')$, $\boldsymbol{\beta} = (\boldsymbol{\gamma}', \boldsymbol{\beta}_0', \boldsymbol{\beta}_1')' : k \times 1$ and $k = k_0 + k_1 + k_2$. An interesting point to observe is that the correlated random effects vector $\mathbf{b}_i$ induces not only intra-cluster correlation but also contemporaneous correlation amongst the latent treatment and potential outcomes. Without any loss of generality, we set the parameter σ_{23} to zero; thus the covariance between the potential outcomes is given by D_{23}. The four free elements of $\boldsymbol{\Sigma}$ are denoted by $\boldsymbol{\sigma} = (\sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{33})$.

The model is completed with a semiparametric distribution on $\lambda_{it} \in R_+$. Specifically, assume that λ_{it} follows an *unknown* probability measure P with distribution G, where G in turn follows a Dirichlet process (DP) prior with base measure G_0 (Fergusson, 1973; Antoniak, 1974). Under this assumption, for any measurable partition $(A_1, A_2, \ldots, A_k)$ of R_+, the random vector $(P(A_1), \ldots, P(A_k))$ is distributed as Dirichlet$(\alpha G_0(A_1), \ldots, \alpha G_0(A_k))$. Formally, the distribution on λ_{it} is formulated as

$$\lambda_{it} \sim G \ (t \leqslant T_i, i \leqslant n),$$

$$G | G_0 \sim \mathrm{DP}(\alpha G_0),$$

$$G_0 = \mathrm{Gamma}\left(\frac{v}{2}, \frac{v}{2}\right), \tag{2}$$

where α is a positive scalar parameter that determines the extent of clustering in the $\{\lambda_{it}\}$. We discuss our choice of α in Section 3 but intuitively α can be viewed as the parameter that measures the prior weight on G_0; larger values of α imply larger prior weight on G_0. Our choice of the base measure G_0 is motivated by the fact that (under G_0) the distribution $\mathbf{z}_{it}|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$ is multivariate-t with location $\mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{b}_i$, dispersion $\boldsymbol{\Sigma}$ and v degrees of freedom. We view this as an appropriate (non-Gaussian) benchmark distribution for $\mathbf{z}_{it}|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

2.2. Model implications

To understand the distributional facets of the proposed model, first observe that the semiparametric structure is induced through the scale parameter λ_{it}. This allows us to move away from normality or from a particular non-Gaussian choice such as the multivariate-t (which is the implied conditional distribution of $\mathbf{z}_{it}$ under G_0). We believe that it is important to specify the distribution flexibly because of the well known fragility of conclusions in treatment models to standard distributional assumptions.

To further understand the model implications, we need two key facts of the Dirichlet process prior on G (Escobar and West, 1995; MacEachern and Muller, 1998). Let

$$\boldsymbol{\lambda} = (\lambda_{11}, \dots, \lambda_{1T_1}, \dots, \lambda_{n1}, \dots, \lambda_{nT_n}) : m \times 1$$

and let $\boldsymbol{\lambda}^{(it)}$ denote the vector $\boldsymbol{\lambda}$ excluding the component λ_{it}. Then, the first fact is that any realization of $\boldsymbol{\lambda}$ from G must lie in a set of $p \leqslant m$ distinct values $\boldsymbol{\phi} = (\phi_1, \dots, \phi_p)$, where the ϕ_i are a random sample from G_0.

From the first fact, it follows that $\boldsymbol{\lambda}^{(it)}$ may contain ties. Suppose that n_{it} values in $\boldsymbol{\lambda}^{(it)}$ are unique. Let

$$(\phi_1^{(it)}, \dots, \phi_{n_{it}}^{(it)})$$

denote those unique values in $\boldsymbol{\lambda}^{(it)}$ and suppose that each unique value $\phi_j^{(it)}$ appears $m_j^{(it)}$ times. Then, the second fact is that the prior distribution of λ_{it} conditioned on $(\boldsymbol{\lambda}^{(it)}, G_0)$, but marginalized over G, can be expressed as

$$\begin{aligned} \lambda_{it}|\boldsymbol{\lambda}^{(it)}, G_0 &\sim \mathrm{E}(G|\boldsymbol{\lambda}^{(it)}, G_0) \\ &\sim \frac{1}{\alpha + m - 1}\,\alpha G_0 + \frac{1}{\alpha + m - 1}\sum_{j=1}^{n_{it}} m_j^{(it)} \delta(\phi_j^{(it)}), \end{aligned} \tag{3}$$

where $\delta(\theta)$ denotes a unit point mass at θ. This is a quite appealing mixture distribution with a continuous component given by the base gamma distribution G_0 and a *random* number of discrete components at the mass-points $\phi_j^{(it)}$.

Let f_N denote the trivariate normal density function. Then, given the above two facts, it follows that the distribution of $\mathbf{z}_{it} = (s_{it}^*, z_{it0}, z_{it1})$ marginalized over λ_{it} and G

$$\mathbf{z}_{it}|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, G_0, \boldsymbol{\lambda}^{(it)} \sim \int f_N(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{b}_i, \lambda_{it}^{-1}\boldsymbol{\Sigma})d[\lambda_{it}|G_0, \boldsymbol{\lambda}^{(it)}]$$

is a mixture distribution with a random number of components. On performing the integration, one gets that $\mathbf{z}_{it}|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, G_0, \boldsymbol{\lambda}^{(it)}$ is distributed as

$$\frac{1}{\alpha+m-1}\,\alpha f_T(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta}+\mathbf{b}_i,\boldsymbol{\Sigma},v)+\frac{1}{\alpha+m-1}\sum_{j=1}^{n_{it}} m_j^{(it)} f_N(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta}+\mathbf{b}_i,\boldsymbol{\Sigma}/\phi_j^{(it)}),$$

where the first component distribution is a multivariate-t distribution with v degrees of freedom and the remaining component distributions are multivariate normal, each with the same mean but different covariance matrix. Taking the product of these mixture distributions, we get that the distribution $f(\mathbf{z}_i|\mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}, G_0, \boldsymbol{\lambda}^{(it)})$ of treatments and potential outcomes over the T_i observations on the ith subject is

$$\prod_{t=1}^{T_i}\left\{\frac{1}{\alpha+m-1}\,\alpha f_T(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta}+\mathbf{b}_i,\boldsymbol{\Sigma},v)+\frac{1}{\alpha+m-1}\right.$$
$$\left.\sum_{j=1}^{n_{it}} m_j^{(it)} f_N(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta}+\mathbf{b}_i,\boldsymbol{\Sigma}/\phi_j^{(it)})\right\}.$$

Although $\mathbf{b}_i$ cannot be marginalized out analytically from this distribution, it is clear that the proposed model induces intra-subject dependence within the context of a general and flexible joint distribution of treatments and outcomes.

3. Prior–posterior analysis

3.1. Estimation by MCMC methods

In this section, we discuss the Bayesian analysis of the proposed model by Markov chain Monte Carlo simulations. These simulations are produced by iterating a Markov chain whose limiting, invariant distribution is the posterior distribution of interest. The sampled variates beyond a transient or burn in stage can, therefore, be viewed as (correlated) draws from the posterior distribution. These sampled draws can be summarized in various ways to produce point and interval estimates of the parameters and posterior density estimates.

Let $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\sigma}, \boldsymbol{\gamma}, \mathbf{D})$ denote the model parameters and let the prior information be represented by the distributions $\boldsymbol{\beta} \sim \mathrm{N}_k(\boldsymbol{\beta}_0, \mathbf{B}_0)$, $\boldsymbol{\sigma} \propto \mathrm{N}_4(\mathbf{g}_0, \mathbf{G}_0)$, restricted to the region that generates a positive definite $\boldsymbol{\Sigma}$ matrix, $\boldsymbol{\gamma} \sim \mathrm{N}_d(\boldsymbol{\gamma}_0, \mathbf{C}_0)$ and $\mathbf{D}^{-1} \sim \mathrm{Wishart}(\rho_0, \mathbf{R}_0)$, where the parameters of the prior, subscripted by zero, are adjusted to represent the available pre-sample information. If the observed data and treatments are denoted by $\mathbf{y} = \{y_{it}\}$ $(t \leqslant T_i, i \leqslant n)$ and $\mathbf{s} = \{s_{it}\}$ $((t \leqslant T_i, i \leqslant n)$, respectively, then the objective is to learn about the posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{s})$ given the data and the prior information.

3.1.1. Posterior sampling

To summarize the unknown posterior density $\pi(\boldsymbol{\theta}|\mathbf{y}, \mathbf{s})$ we augment the parameter space (following Tanner and Wong (1987), Chib (1992) and Albert and Chib

(1993)) to include $\{\mathbf{z}_{it}\}$ and $\{\lambda_{it}\}$ for $(t \leqslant T_i, i \leqslant n)$ and $\{\mathbf{b}_i\}$ for $(i \leqslant n)$, where $\mathbf{z}_{it} = (s^*_{it}, z_{it0}, z_{it1})$ and focus on the sampling of the distribution $\pi(\boldsymbol{\theta}, \{\mathbf{z}_{it}\}, \{\mathbf{b}_i\}, \{\lambda_{it}\} | \mathbf{y}, \mathbf{s})$. To cope with the high dimension of the target distribution, and to promote mixing of the Markov chain output, the vector $\boldsymbol{\sigma}$ is sampled marginalized over $\{\mathbf{z}_{it}\}$, and the vector $\boldsymbol{\beta}$ marginalized over $\{\mathbf{b}_i\}$.

MCMC algorithm for sampling $\pi(\boldsymbol{\theta}, \{\mathbf{z}_{it}\}, \{\mathbf{b}_i\}, \{\lambda_{it}\} | \mathbf{y}, \mathbf{s})$

1. *Sample* $(\boldsymbol{\sigma}, \{\mathbf{z}_{it}\})$ *from* $\boldsymbol{\sigma}, \{\mathbf{z}_{it}\} | \mathbf{y}, \mathbf{s}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \boldsymbol{\beta}$ *by drawing*
(a) $\boldsymbol{\sigma}$ *from* $\boldsymbol{\sigma}, | \mathbf{y}, \mathbf{s}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \boldsymbol{\beta}$ *and*
(b) $\mathbf{z}_{it}$ *from* $\mathbf{z}_{it} | s_i, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \boldsymbol{\Sigma}, \{\lambda_{it}\}$, independently for $i = 1, \ldots, n$ and $t \leqslant T_i$;
2. *Sample* $(\boldsymbol{\beta}, \{\mathbf{b}_i\})$ *from the distribution* $(\boldsymbol{\beta}, \{\mathbf{b}_i\}) | \mathbf{y}, \mathbf{s}, \{\mathbf{z}_{it}\}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \mathbf{D}, \boldsymbol{\Sigma}$ *by drawing*
(a) $\boldsymbol{\beta}$ *from* $\boldsymbol{\beta} | \{\mathbf{z}_{it}\}, \mathbf{D}, \boldsymbol{\Sigma}, \{\lambda_{it}\}$ *and*
(b) $\{\mathbf{b}_i\}$ *from* $\mathbf{b}_i | \{\mathbf{z}_{it}\}, \boldsymbol{\beta}, \mathbf{D}, \{\lambda_{it}\}$;
3. *Sample* γ *from* $\gamma | \{\mathbf{b}_i\}, \mathbf{D}^{-1}$;
4. *Sample* $\mathbf{D}^{-1}$ *from* $\mathbf{D}^{-1} | \{\mathbf{b}_i\}, \boldsymbol{\gamma}$;
5. *Sample* $\{\lambda_{it}\}$ *from* $\lambda_{it} | \mathbf{z}_{it}, \mathbf{b}_i, \boldsymbol{\lambda}^{(it)}, G_0, \boldsymbol{\beta}, \boldsymbol{\Sigma}$;
6. *Repeat Steps* 1–5 *using the most recent values of the conditioning variables.*

We now touch on the main ideas behind this algorithm, deferring full details to Appendix A. In Step 1(a) of this algorithm, $\boldsymbol{\sigma}$ is sampled from the distribution

$$\pi(\boldsymbol{\sigma} | \mathbf{y}, \mathbf{s}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \boldsymbol{\beta}) \propto \pi(\boldsymbol{\sigma}) \prod_{i=1}^{n} \prod_{t=1}^{T_i} f(y_{it}, s_{it} | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}), \tag{4}$$

where, on letting Φ denote the cdf of the standard normal density function,

$$\begin{aligned} f(y_{it}, s_{it} = 0 | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) &= f(y_{it} | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(s_{it} = 0 | y_{it}, \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= f_N(y_{it} | \mathbf{x}'_{it0} \boldsymbol{\beta}_0 + b_{i0}, \lambda_{it}^{-1} \sigma_{22}) \\ &\quad \times \Phi \left(\frac{-\mathbf{w}'_{it} \boldsymbol{\gamma} - a_i - \sigma_{12} \sigma_{22}^{-1} (y_{it} - \mathbf{x}'_{it0} \boldsymbol{\beta}_0 - b_{i0})}{(1 - \sigma_{12}^2 (\lambda_{it} \sigma_{22})^{-1})^{1/2}} \right) \end{aligned} \tag{5}$$

and

$$\begin{aligned} f(y_{it}, s_{it} = 1 | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) &= f(y_{it} | \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) p(s_{it} = 1 | y_{it}, \mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \\ &= f_N(y_{it} | \mathbf{x}'_{it1} \boldsymbol{\beta}_1 + b_{i1}, \lambda_{it}^{-1} \sigma_{33}) \\ &\quad \times \Phi \left(\frac{\mathbf{w}'_{it} \boldsymbol{\gamma} + a_i + \sigma_{13} \sigma_{33}^{-1} (y_{it} - \mathbf{x}'_{it1} \boldsymbol{\beta}_1 - b_{i1})}{(1 - \sigma_{13}^2 (\lambda_{it} \sigma_{33})^{-1})^{1/2}} \right). \end{aligned} \tag{6}$$

It is clear that this posterior density does not belong to a known family of distributions but the sampling strategy that is developed by Chib and Greenberg (1998) in a related context can be used. Essentially, the idea is to employ the Metropolis–Hastings algorithm with a proposal density that is matched to the target density around the mode (Chib and Greenberg, 1995). Details are given in Appendix A.

In Step 1(b), the latent values $\mathbf{z}_{it}$ are sampled from $\mathbf{z}_{it}|s_{it}, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \boldsymbol{\Sigma}, \{\lambda_{it}\}$ based on Chib (1992) and Albert and Chib (1993). One point to stress is that for any given unit, either z_{it1} is drawn or z_{it0} is drawn, but not both. In Step 2(a) of the algorithm the sampling of the coefficients $\boldsymbol{\beta}$ is done marginalized over $\{\mathbf{b}_i\}$ because Chib and Carlin (1999) have shown that this strategy reduces the serial correlation of the MCMC chain. From the Bayes theorem, the desired posterior distribution of $\boldsymbol{\beta}$, marginalized over the random effects, is

$$\pi(\boldsymbol{\beta}|\{\mathbf{z}_{it}\}, \{\lambda_{it}\}, \boldsymbol{\gamma}, \mathbf{D}, \boldsymbol{\Sigma}) \propto \pi(\boldsymbol{\beta}) \prod_{i=1}^{n} f(\mathbf{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{D}, \boldsymbol{\Sigma}), \tag{7}$$

where $\mathbf{z}_i = (z_{i1}, \ldots, z_{iT_i})$ denotes the $3T_i \times 1$ vector of observations on the ith subject and

$$\begin{aligned} f(\mathbf{z}_i|\boldsymbol{\lambda}_i, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{D}, \boldsymbol{\Sigma}) &= \int f(\mathbf{z}_i|\mathbf{b}_i, \boldsymbol{\lambda}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma})\pi(\mathbf{b}_i|\boldsymbol{\gamma}, \mathbf{D})\,\mathrm{d}\mathbf{b}_i \\ &\propto \int \exp\{-0.5(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta} - r\mathbf{W}_i\mathbf{b}_i)'\boldsymbol{\Sigma}_i(\mathbf{z}_i - \mathbf{X}_i\boldsymbol{\beta} - \mathbf{W}_i\mathbf{b}_i)\} \\ &\quad \times f_N(\mathbf{b}_i|\mathbf{V}_i\boldsymbol{\gamma}, \mathbf{D})\,\mathrm{d}\mathbf{b}_i, \end{aligned}$$

$\mathbf{W}_i = (\mathbf{I}_3, \mathbf{I}_3, \ldots, \mathbf{I}_3)' : 3T_i \times 1$, $\boldsymbol{\Sigma}_i = \boldsymbol{\Lambda}_i^{-1} \otimes \boldsymbol{\Sigma}$ and $\boldsymbol{\Lambda}_i^{-1} = \mathrm{diag}\,(\lambda_{i1}^{-1}, \ldots, \lambda_{iT_i}^{-1})$. After some algebra the density in (7) is seen to be Gaussian with mean $\hat{\boldsymbol{\beta}} = \mathbf{B}(\boldsymbol{\beta}_0\mathbf{B}_0^{-1} + \sum_{i=1}^{n} \mathbf{X}_i'\boldsymbol{\Omega}_i^{-1}$ $(\mathbf{z}_i - \mathbf{W}_i\mathbf{V}_i\boldsymbol{\gamma}))$ and variance $\mathbf{B} = (\mathbf{B}_0^{-1} + \sum_{i=1}^{n} \mathbf{X}_i\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i)^{-1}$, where $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i' + \boldsymbol{\Sigma}_i$.

The next part of the MCMC algorithm requires the sampling of the random effects $\{\mathbf{b}_i\}$, the random effects coefficients $\boldsymbol{\gamma}$ and random effects variance $\mathbf{D}^{-1}$. The distributions that need to be sampled in each of these cases follow from standard results for Bayesian longitudinal models and are presented in Appendix A.

Finally, in Step 5 $\{\lambda_{it}\}$ is sampled from $\lambda_{it}|\mathbf{z}_{it}, \mathbf{b}_i, \boldsymbol{\lambda}^{(it)}, G_0, \boldsymbol{\beta}, \boldsymbol{\Sigma}$. Under the distribution of λ_{it} given in (3), a direct calculation shows that the updated distribution is a continuous-discrete distribution.

$$\lambda_{it}|\mathbf{z}_{it}, \mathbf{b}_i, \boldsymbol{\lambda}^{(it)}, G_0, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim q_{it0}\pi_0(\lambda_{it}|\mathbf{z}_{it}, \boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma}) + \sum_{j=1}^{p_{it}} q_{itj}\delta(\phi_j^{(it)}),$$

$$t = 1, \ldots, T_i; \quad i = 1, \ldots, n, \tag{8}$$

where

$$\begin{aligned} \pi_0(\lambda_{it}|\mathbf{z}_{it}, \boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma}) &\propto f(\mathbf{z}_{it}|\boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma})dG_0(\lambda_{it}) \\ &\propto \lambda_{it}^{(v+3)/2-1}\mathrm{e}^{-\lambda_{it}d_{it}} \end{aligned}$$

is a gamma density with parameters $(v+3)/2$ and $d_{it} = (v + (\mathbf{z}_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - \mathbf{b}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - \mathbf{b}_i))/2$ and the weights q_{it0} and q_{itj} are defined in Appendix A. Each λ_{it} $(t = 1, \ldots, T_i;\ i = 1, \ldots, n)$ is sampled from these mixed distributions given the most current value of $\boldsymbol{\lambda}^{(it)}$.

3.2. Treatment effects and Bayesian matching

Given the observed data s_{it} and y_{it}, define the *unit* (subject-time) level treatment effect δ_{it} as

$$\delta_{it} = \begin{cases} z_{it1} - y_{it} & \text{if } s_{it} = 0, \\ y_{it} - z_{it0} & \text{if } s_{it} = 1. \end{cases} \tag{9}$$

The goal is to understand how this treatment effect may be calculated. It is easy to see that δ_{it} is not non-parametrically identified because it requires knowledge of the counter-factual which is never observed. Rosenbaum and Rubin (1983), working in the cross-section case, have provided one way to deal with the identifiability issue. Suppose one assumes that

$$(z_{it0}, z_{it1}) \perp s_{it} | w_{it},$$

where the subscript t is retained to avoid introducing new notation. If $p(w_{it}) = \Pr(s_{it} = 1|w_{it})$ denotes the probability of treatment (the propensity score), then Rosenbaum and Rubin (1983) show that one can identify the expected value of $\delta_{it} = y_{it} - z_{it0}$ given the treatment (the "treatment effect for the treated") as

$$\begin{aligned} \mathrm{E}(\delta_{it}|s_{it} = 1) &= \mathrm{E}(y_{it} - z_{it0}|s_{it} = 1) \\ &= \mathrm{E}^{w_{it}|s_{it}=1}\mathrm{E}(y_{it} - z_{it0}|s_{it} = 1, w_{it}) \\ &= \mathrm{E}^{w_{it}|s_{it}=1}\{\mathrm{E}(y_{it}|s_{it} = 1, w_{it}) - \mathrm{E}(z_{it0}|s_{it} = 0, w_{it})\} \\ &= \mathrm{E}^{w_{it}|s_{it}=1}\{\mathrm{E}(y_{it}|s_{it} = 1, p(w_{it})) - \mathrm{E}(z_{it0}|s_{it} = 0, p(w_{it}))\}, \end{aligned}$$

where the second line utilizes the law of the iterated expectation, the third line that $\mathrm{E}(z_{it0}|s_{it} = 1, w_{it}) = \mathrm{E}(z_{it0}|s_{it} = 0, w_{it})$ from the assumed conditional independence assumption and the fourth line that $\mathrm{E}(z_{it0}|s_{it} = 0, w_{it}) = \mathrm{E}(z_{it0}|s_{it} = 0, x_{it}, p(w_{it}))$. Each of the two terms in braces in the fourth line can be estimated by forming matched treated and untreated groups, matched according to the propensity score, and then computing the difference in average outcomes for the respective groups (it is important to note that to estimate the expectations efficiently, it is necessary to adjust for the influence of covariates, say by modeling the outcomes by a regression; thus, even this approach requires a covariate model for both the treatments and outcomes). The outer expectation is estimated by the average of these differences across different values of the propensity score.

Some difficulties arise if this approach is applied in the longitudinal context. If we let z_{i0} and z_{i1} be vectors of observations on the ith subject and let s_i denote the $T_i \times 1$ sequence of treatments, then the assumption that $(z_{i0}, z_{i1}) \perp s_i | \{w_{it}\}$ is difficult to sustain given the clustering in treatment and outcomes that is likely to be present in longitudinal data. In fact, one may question whether the conditional independence assumption can be satisfied in practice, even with cross-section data, given that unobserved variables which jointly affect the treatment and the outcomes are the norm rather than the exception in most applications.

An alternative approach is to suppose that there is an additional source of information—a valid randomly assigned covariate (instrument)—and to model the joint distribution of treatments and potential outcomes in a distributionally flexible way. The effect of covariates is modeled by making functional form assumptions. There seems to be no simple way to relax the latter feature given that in social-science applications both treatments and outcomes are typically affected by a large number of covariates. As noted above, even in propensity score approaches, the effect of covariates must be modeled parametrically in both the treatment and outcome distributions. Thus, it seems to us, that provided one can justify the exclusion restriction in any given empirical problem, our approach provides a useful framework for tackling the treatment effect problem with longitudinal data.

We now briefly discuss how one can use our model to obtain summaries of the treatment effects, including one that is based on matched groups, conditioned on the observed data, but marginalized *over all unknowns* including the random effects and the unknown distribution G.

The goal is to derive various treatment effect distributions from the posterior distribution of δ_{it} given the data (y, s). This conditioning on the observed data, which is quite natural in the Bayesian context, is rarely used when calculating treatment effects. By definition, the posterior density of δ_{it} is

$$\pi(\delta_{it}|\mathbf{y},\mathbf{s}) = \int \pi(\delta_{it}|\mathbf{y},\mathbf{s},\boldsymbol{\theta},s^*_{it},\lambda_{it},\mathbf{b}_i)\,d\pi(\boldsymbol{\theta},s^*_{it},\lambda_{it},\mathbf{b}_i|\mathbf{y},\mathbf{s}), \tag{10}$$

from which a sample of δ_{it} can be produced by the method of composition using the draws z_{it1} or z_{it0} from Step 1(d) of the MCMC algorithm in Appendix A. The calculation of these posterior distributions is akin to the way in which Albert and Chib (1995) find the posterior distribution of latent residuals in binary data models. Given a posterior sample of δ_{it} from $\pi(\delta_{it}|y,s)$, which we denote by $\{\delta_{it}^{(g)}\}$, the *unit mean effect* $\bar{\delta}_{it} = \mathrm{E}(\delta_{it}|y,s)$ may be estimated as

$$\bar{\delta}_{it} \approx G^{-1}\sum_{g=1}^{G}\delta_{it}^{(g)}.$$

In practice, it is useful to consider treatment effects at a more aggregated level. For example, one can define the treatment effect at the subject level $\delta_i = T_i^{-1}\sum_{t=1}^{T_i}\delta_{it}$. A posterior sample on δ_i is available as $\{T_i^{-1}\sum_{t=1}^{T_i}\delta_{it}^{(g)}\}_{g=1}^{G}$ from which one can calculate the *subject level mean treatment effect* $\bar{\delta}_i$ where

$$\bar{\delta}_i \approx G^{-1}\sum_{g=1}^{G}T_i^{-1}\sum_{t=1}^{T_i}\delta_{it}^{(g)}$$

$$= T_i^{-1}\sum_{t=1}^{T_i}\bar{\delta}_{it}.$$

The posterior standard deviation of δ_i can be estimated as the sample standard deviation of the draws on δ_i. Similarly, the treatment effect for a randomly selected observation from the population may be defined as

$$\delta = \frac{\sum_{i,t} \delta_{it}}{m} = \frac{\sum_i \delta_i}{n}$$

whose posterior distribution is again available from the posterior sample on δ_{it}. The mean of the posterior distribution of δ may be called the *average mean treatment effect*. Another useful aggregate quantity, following Heckman and Robb (1985), is the treatment effect for the treated. Let $T^* = \{i, t : s_{it} = 1\}$ denote the set of units that receive the treatment and consider the posterior distribution of δ_{it} for units that are in T^*. A relevant summary of these posterior distributions is the *average mean treatment effect for the treated*

$$\delta^* = n^{*-1} \sum_{i,t \in T^*} \mathrm{E}(\delta_{it} | \mathbf{y}, \mathbf{s})$$

$$\approx n^{*-1} \sum_{i,t \in T^*} G^{-1} \sum_{t=1}^{T_i} \delta_{it}^{(g)}, \tag{11}$$

where n^* is the cardinality of T^*.

Finally, consider the treatment effect across units, grouped not by subject or treatment status, but according to the probability of treatment. Let

$$p_{it} = \Phi(\mathbf{w}_{it}' \boldsymbol{\gamma} + a_i) \tag{12}$$

denote the unit level *treatment probability* given the covariates and the random effect. This treatment probability is distinct from the propensity score that is defined in the standard matching framework in that it includes the unobservable random effect. Now suppose that the range of p_{it} is divided into 10 equally spaced intervals (other intervals can be treated in the same way). Let

$$D_j = \left\{ i, t : p_{it} \in \left(\frac{j-1}{10}, \frac{j}{10} \right) \right\}, \quad j = 1, 2, \ldots, 10,$$

denote the set of units for which the corresponding treatment probability is in the jth decile. Units that fall into any one of these decile groups have similar, or matched, treatment probabilities. There are two important points to note about this grouping that distinguish it from the classical matching described above. First, this matching occurs on the basis of both the observed covariates w_i and unobserved covariate a_i. Second, even though the decile groups are set at the outset, the posterior distribution of the treatment effect is well defined even in the low and high decile groups where most units are likely to have the same observed treatment status. Units in those extreme deciles are essentially self-matched. Given the grouped units, define the matched treatment effect

$$\delta_j = n_j^{-1} \sum_i \sum_t \delta_{it}, (i, t) \in D_j, \quad j = 0, 1, \ldots, 10,$$

where n_j denotes the number of units in D_j. The posterior distribution of δ_j can be derived from the posterior sample of δ_{it} as follows. Specifically, given a draw of the parameters γ and random effects a_i at the gth MCMC iteration, one calculates the unit level treatment probability $p_{it}^{(g)} = \Phi(\mathbf{w}_{it}'\gamma^{(g)} + a_i^{(g)})$. Each unit, along with the associated value of $\delta_{it}^{(g)}$, is then placed in the appropriate decile group and the matched treatment effects are averaged. This constitutes a draw from the posterior distribution of δ_j. This process of matching and averaging within each decile group is repeated for each MCMC iteration and the sample $\{\delta_j^{(g)}\}$ thus created is summarized by the group-specific averages $G^{-1}\sum \delta_j^{(g)}$. These averages are the Bayesian matched treatment effect for the jth decile grouping. As we show in the example that follows, the δ_j provides a convenient way to isolate the heterogeneity in treatment effects, by treatment probability.

4. Example: unions and the wage premium

We now present a real data example to highlight the variety of treatment effect distributions which may be constructed in our framework. Our example is concerned with the wage premium associated with union membership. In this problem, the treatment variable s_{it} is one if the subject is a union member in year t, and zero if not. The unit level treatment effect δ_{it} is the difference in the logarithm of the hourly wage in the union and non-union sectors for subject i in year t. Studies dealing with the union wage premium abound in the literature (for example, see Lewis (1986) who isolates over 200 articles dealing with this problem and reports union wage premia ranging from −75% to 95%). In this literature, the non-random sorting of individuals to the two sectors is central to the debate. Some studies (e.g., Freeman, 1984; Lewis, 1986) argue that unionized firms are able to select more able workers. Simple OLS estimates of the union wage premium will, therefore, provide an upper bound of the true union wage effect, since more productive workers will tend to join the union. Other studies, such as Robinson (1989), claim that union pay scales are less sensitive to individual ability, implying that less productive workers are more likely to join the union. In this case, OLS estimates provide a lower bound of the true effect. Robinson (1989) also argues that the sector of relative advantage may be different for different workers. Some workers may do better in the union sector, while others will be more successful in the non-union sector. Consequently, accounting for heterogeneity and endogeneity of the treatment is important for understanding the union wage premium.

To examine the relationship between unions and wages, a random sample of 241 white male high school graduates was drawn from the National Longitudinal Survey of Youth (NLSY) covering the period 1982–1991. Individuals who had not completed their schooling, or who had dropped out of the NLSY at some point during this period were excluded from the analysis. In our sample, the yearly unionization rates range from 24% to 32%. Overall, 60% of the subjects were in a union job at some point during the sample period.

In our model we set $x_{it0} = x_{it1}$ and let x_{it0} consist of the variables labor market experience and its square, marital status, the unemployment rate in the county of residence at time t, and a linear time trend. The covariate vector w_{it} includes x_{it0} and a set of variables excluded from the wage equations. These covariates are an indicator

Table 1
Union data example: summary of means and standard deviations of variables

Variable	All	$s_{it} = 1$[a]	$s_{it} = 0$
Hourly pay (real $)	6.759 (2.961)	8.149 (2.949)	6.181 (2.768)
Experience (in years)	9.176 (3.584)	9.946 (3.456)	8.856 (3.589)
Married (0–1)	0.579 (0.494)	0.702 (0.458)	0.528 (0.499)
Unemployment rate	0.083 (0.033)	0.084 (0.033)	0.083 (0.033)
Professional spouse (0–1)	0.101 (0.302)	0.113 (0.317)	0.096 (0.295)
% employed in Manuf	0.262 (0.104)	0.274 (0.100)	0.257 (0.105)
Right to work state	0.177 (0.382)	0.133 (0.340)	0.196 (0.397)
State UI takeup rate	0.668 (0.154)	0.673 (0.151)	0.666 (0.156)
State max UI benefit ($)	188 (43)	190 (45)	187 (42)
State tuition ($)	1641 (699)	1721 (705)	1608 (694)

[a] The treatment variable $s_{it} = 1$ represents union membership.

Table 2
Union data example: estimates of the union wage premium using standard methods in the literature

Method	Estimated union wage premium
Pooled OLS	0.244 (0.017)
Pooled IV	0.204 (0.172)
Panel data fixed effects	0.148 (0.017)
Panel data random effects	0.161 (0.016)
Panel data IV	0.173 (0.212)

Estimate is the coefficient on union status variable in a regression of y_{it} on x_{it} and union status. Instruments are the w_{it} covariate vector.

of whether the subject's spouse has a professional or managerial job (such families may be more likely to already have the benefits that unions offer) and the fraction of the local labor force employed in manufacturing (a proxy for local union domination). Following Budd and Na (2000), we also include data on whether the subject resides in a right to work state; the unemployment insurance takeup rate in the state at time t; the maximum unemployment insurance benefits; and public university tuition in the subject's state of residence. The last three variables capture the attractiveness of union non-wage benefits.

A summary of the data used in the analysis is given in Table 1. Union members earn almost two dollars more per hour in the sample, although non-union members tend to have less experience and are less likely to be married. In order to provide some context, we present in Table 2 estimates of the union wage premium using some of the estimation methods discussed in the literature. From the first row of the table we see that a simple pooled OLS technique produces a wage of premium of approximately 24%. Accounting for the possible endogeneity of union status using pooled instrumental variable methods reduces the union premium to 20%, which is no longer significant. If we account for subject-level differences in the wage outcomes, then the estimated union wage premium falls to approximately 16%. None of these methods, however, are capable of isolating heterogeneity in the treatment effect.

4.1. Model fitting

To analyze these data, we employ the general model discussed above, set v to 10 and the off-diagonal terms of $\mathbf{\Sigma}$ to zero. The treatment and potential outcomes remain correlated at each point in time because of the correlated random-effect vector $\mathbf{b}_i$. For simplicity, we do not model time series dynamics in treatments and outcomes although it is possible that a full analysis of these data may require such extended modeling. This is because our main (and limited) goal is to illustrate the workings of our proposed methods. If necessary, further realism can be introduced along the lines discussed by Geweke and Keane (2000), Hirano (1999), and Vella and Verbeek (1998) in other contexts.

Because our model contains a large number of parameters and prior elicitation is difficult, we build our prior distribution by analyzing a prior, training sample data set. The training sample is formed by randomly selecting 60 subjects from the National Longitudinal Survey of Youth. The data on these individuals are then subjected to analysis with default priors on the parameters. The training sample posterior mean is taken to be mean of our prior distribution and the variance of our prior distribution is taken to be a multiple of the training sample posterior covariance. It should be noted that because we adjust the covariance matrix from the training sample, and use specific distributional forms to represent our prior information, this training sample approach is not tantamount to analyzing the whole sample with a default prior on the parameters. The training sample approach may be viewed as replicating the usual Bayesian sequential logic, where a previous analysis with a different data set is used as the basis for prior formulation for the current problem. We summarize the results from the training sample in Table 3 under the heading "training sample posterior."

To analyze the remaining data, we set the smoothness parameter α to yield about 160 –185 clusters at each iteration of the sampler. We run our sampler for 10,000 cycles with a burn-in of 1000 cycles. This MCMC sample size design appears to be adequate since the autocorrelation plots in Fig. 1 indicate that our sampler mixes relatively well.

We summarize the results from the fitting in Table 3. One can see from the posterior estimates that a person living in a right to work state has a substantially lower probability of being a union member. From the parameter estimates of experience and experience squared in the potential outcome (log wage) equations, we see that non-union members have steeper wage profiles. The positive estimate of D_{12} suggests that unobservable factors affecting union choice are correlated with factors that influence the potential outcomes in the non-union sector. Union members appear to have better non-union alternatives than workers who are actually employed in the non-union sector. The estimated positive value of D_{23} indicates that individuals with high earnings at non-union jobs tend to have high earnings at unionized jobs, suggesting that sectoral abilities are positively related but not perfectly correlated.

Table 3
Union data example: posterior estimates based on data for the period 1982–1991

Variable	Training sample posterior	Posterior
Treatment (Union member)		
Intercept	−7.556 (1.887)	−4.504 (1.033)
Experience	1.014 (0.266)	0.402 (0.138)
Experience-squared/10	−0.256 (0.073)	−0.087 (0.048)
Married	0.060 (0.255)	0.549 (0.179)
Unemployment rate	0.111 (0.048)	−0.003 (0.032)
Time trend	−0.456 (0.231)	−0.218 (0.112)
Spouse prof.	0.075 (0.409)	0.165 (0.209)
% Manufacturing	4.051 (2.785)	1.110 (1.472)
Right to work state	−1.965 (1.053)	−0.970 (0.401)
UI takeup	0.666 (0.807)	−0.087 (0.442)
Max UI benefit	−0.006 (0.006)	−0.003 (0.004)
Tuition	−0.0003 (0.001)	0.001 (0.0007)
Potential outcomes (z_{it0})		
Intercept	1.028 (0.414)	1.130 (0.116)
Experience	0.061 (0.074)	0.105 (0.019)
Experience-squared	−0.007 (0.014)	−0.034 (0.004)
Married	0.086 (0.064)	0.021 (0.017)
Unemployment rate	−0.0001 (0.008)	0.002 (0.003)
Time trend	−0.021 (0.065)	−0.016 (0.017)
Potential outcomes (z_{it1})		
Intercept	1.328 (0.256)	1.268 (0.064)
Experience	0.111 (0.048)	0.087 (0.012)
Experience-squared	−0.036 (0.006)	−0.020 (0.003)
Married	−0.009 (0.027)	0.021 (0.014)
Unemployment rate	−0.012 (0.004)	−0.009 (0.002)
Time trend	−0.017 (0.046)	−0.019 (0.011)
σ_{22}	0.028 (0.005)	0.007 (0.001)
σ_{33}	0.034 (0.005)	0.018 (0.002)
D_{12}	1.323 (0.607)	0.846 (0.252)
D_{13}	−0.033 (0.309)	0.020 (0.075)
D_{22}	0.903 (0.211)	0.228 (0.048)
D_{33}	0.468 (0.091)	0.105 (0.011)
D_{23}	0.039 (0.087)	0.052 (0.013)
Treatment effects		
δ		0.005 (0.068)
δ^* (union)		0.272 (0.028)
δ^* (non-union)		−0.102 (0.096)

The second column gives the posterior means and standard deviations from the training sample. The corresponding results for the analysis sample are in the last column. Results are based on 10,000 MCMC draws.

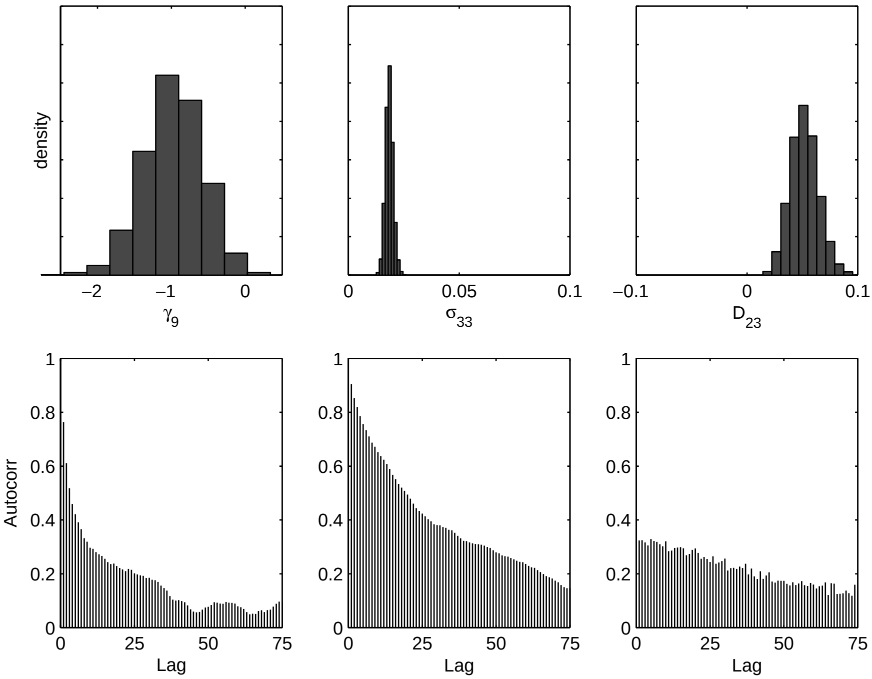

Fig. 1. Union data example: posterior distributions of γ_9, σ_{33}, and D_{23} (top panel) and corresponding autocorrelation plots from the MCMC output (bottom panel).

The bottom three rows of Table 3 contain the posterior means and standard deviations of the average mean treatment effect, δ, as well as the treatment effect for union and non-union members. We see that our model, which accounts for the endogeneity of union status, produces a substantially lower estimate of the treatment effect than a model that ignores the endogeneity of the treatment. Our model also generates a smaller union wage gap than that obtained from the IV fit reported in Table 2. However, the estimates of the treatment effect for the treated, δ^*, suggest substantial heterogeneity in the union wage premium. This is not captured by standard IV approaches. Finally, we estimated parametric specifications of our model assuming multivariate normal and multivariate-t (with 10 degrees of freedom) distributions. These parametric models yield estimates of δ that lie between the semiparametric and IV estimates (e.g., $\delta = 0.05$ or 0.06). The estimate of D_{12} appears to be particularly sensitive to the parametric assumption (the mean posterior of D_{12} was 0.499 in the multivariate-t model, compared to 0.846 in the semiparametric model), leading to an estimated treatment effect of -0.030 for non-union workers, as compared to $\delta^* = -0.102$ from the semiparametric model. Experiments by the authors using simulated data suggest that our semiparametric model is particularly robust to extreme observations, which may explain these differences in results.

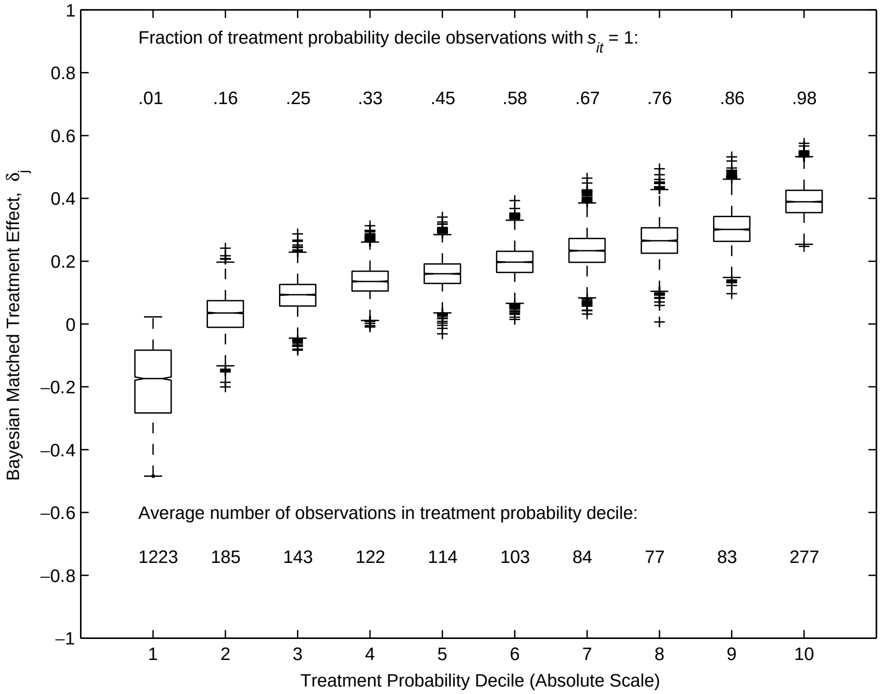

Fig. 2. Union data example: box plots of δ_j stratified by treatment probability decile. The mean number of observations falling into the treatment probability decile is reported at the bottom of the graph, and the fraction of these observations that receive the treatment is reported at the top.

4.2. Treatment distributions

Fig. 2 presents the boxplots of the posterior distributions of the Bayesian matched treatment effects, $\delta_1,\ldots,\delta_{10}$. This figure is useful in understanding the heterogeneity in the treatment effect across the various treatment probability deciles. Workers appear to self-sort across sectors, since the distribution of δ_j becomes more positive for the higher treatment probability deciles (i.e., observations with a higher probability of union membership). Individuals in the lowest treatment probability decile, for whom the predicted probability of being in a union job is between 0 and 0.1, are estimated to earn 18% more, on average, in the non-union sector. The union wage premium is estimated to be positive for the remaining set of workers; the union wage-gap is estimated to be 18% or more for the five highest treatment probability deciles. These findings are consistent with a model of the labor market in which workers choose the sector in which they have a comparative advantage. The numbers at the top and bottom of the figure indicate the average number of observations in the sample that fall into a particular treatment probability decile, and the probability that an observation received the treatment (i.e., held a union job).

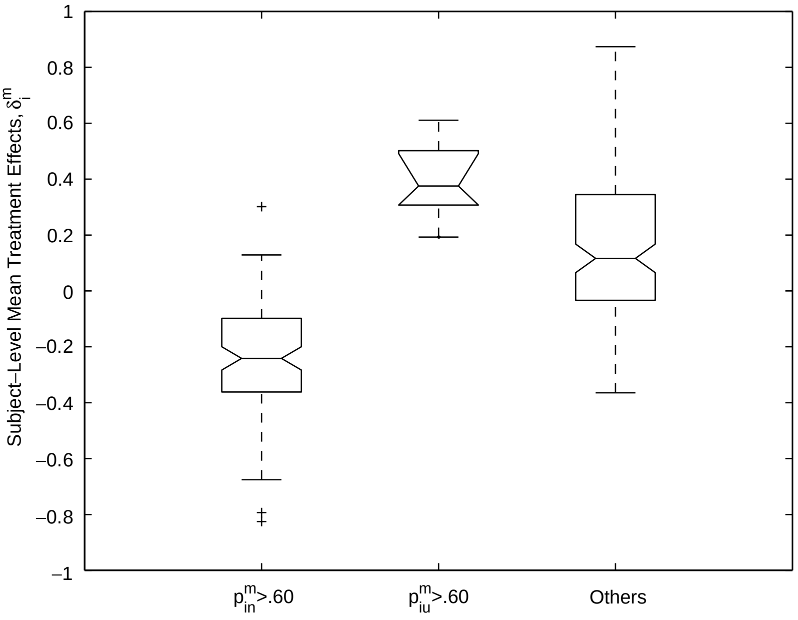

Fig. 3. Union data example: box-plots of $\bar{\delta}_i$ categorized by $\bar{p}_{iu}$ (left), $\bar{p}_{in}$ (middle) and others (right).

For example, only 1% of the estimated 1223 observations with treatment probability between 0 and 0.1 received the treatment. As might be expected, the δ_1 distribution exhibits more variation than the treatment distributions in the other deciles because the observations in the first decile largely contain individuals who are self-matched.

Angrist et al. (1996) and Heckman et al. (1998) argue that some subjects in a sample are not at risk of changing treatments, and so the data are not very informative about their treatment effect distributions. One way to capture this idea in our model is to group subjects by their predicted probabilities of receiving the treatment (being a union member) in every period in the sample, denoted by $\bar{p}_{iu}$, and never receiving the treatment in the sample, $\bar{p}_{in}$. Subjects with larger values of $\bar{p}_{iu}$ and $\bar{p}_{in}$ are unlikely to be at risk of changing treatment. In this example, we considered subjects with $\bar{p}_{iu} > 0.60$ to be likely to always work at a union job, and those with $\bar{p}_{in} > 0.60$ to be very likely to never be a union member. Fig. 3 shows that the empirical distribution of the subject level mean treatment effect, $\bar{\delta}_i$, for subjects presumed to be at risk of changing treatment lies between that for subjects with high values of $\bar{p}_{iu}$ and $\bar{p}_{in}$. One interpretation of this figure is that if employment in a union job were to become more attractive, one would likely see a small increase in wages, since the subjects induced to change sectors would be drawn from the sub-group labeled “others” in Fig. 3. For these subjects, the treatment effect distribution is centered at approximately 0.10 and crosses zero.

5. Conclusion

This paper has developed a semiparametric Bayesian analysis of treatment models for longitudinal data that incorporates potential outcomes and non-random treatment assignment. The model incorporates the special features of longitudinal data, for example subject-specific clustering of treatments and outcomes, without requiring independence of treatments and responses given the covariates or strong distributional assumptions. It is important to specify the distribution flexibly because of the well known fragility of conclusions in treatment models to usual distributional assumptions.

Unlike other semiparametric approaches, the unit level treatment effects generated from the proposed model may be summarized to yield many of the effects discussed in the literature. For example, we construct the treatment effect for the treated, as well as the average treatment effect in the population. Furthermore, by relying on a model-based formulation of the outcomes, we obtain detailed summaries of the treatment effects, including one that is based on matched groups, conditioned on the observed data, but marginalized over all unknowns in the model. The empirical example also emphasizes that our approach provides an intuitive and convenient way of summarizing treatment effect heterogeneity. We should mention that this paper does not delve into the problem of comparing the proposed model with parametric alternatives. Such a comparison is now possible for the first time using the methods that have been developed by Basu and Chib (2001). An application of the latter methods to the current problem will be considered in future work.

We conclude by noting that the proposed framework may be utilized in situations that are more general than the binary treatment, continuous outcome case considered here. More complicated time series dynamics may also be considered in the treatment and outcome equations. Finally, it is possible to allow for multiple treatments or multivariate responses, as well as treatments and outcomes which are binary, ordinal, or continuous.

Acknowledgements

We thank John Budd for proving the state level data, and the editor, associate editor and four reviewers for their helpful comments.

Appendix A

In this appendix we summarize the MCMC algorithm that is used to fit the semiparametric panel potential outcomes model proposed in this paper.

Algorithm.

1. Sample $(\boldsymbol{\sigma}, \{\mathbf{z}_{it}\})$ from $\boldsymbol{\sigma}, \{\mathbf{z}_{it}\}|\mathbf{y}, \mathbf{s}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \boldsymbol{\beta}$ by drawing

(a) $\boldsymbol{\sigma}$ from $\boldsymbol{\sigma}, |\mathbf{y}, \mathbf{s}, \{\mathbf{b}_i\}, \{\lambda_{it}\}, \boldsymbol{\beta}$ which is proportional to

$$g(\boldsymbol{\sigma}) = \pi(\boldsymbol{\sigma})\prod_{i=1}^{n}\prod_{t=1}^{n_i} f(y_{it}, s_{it}|\mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$$

To sample $g(\boldsymbol{\sigma})$, following Chib and Greenberg (1998), let $q(\boldsymbol{\sigma}|\boldsymbol{\mu}, \tau\mathbf{V})$ denote a multivariate-t density with parameters $\boldsymbol{\mu}$ and $\mathbf{V}$ defined as the mode and inverse of the negative Hessian, respectively, of $\log g(\boldsymbol{\sigma})$, where τ is an additional scaler tuning parameter. Let the degrees of freedom of this density be fixed at a value (say 15). Then

(b) Sample a proposal value $\boldsymbol{\sigma}'$ from the density $q(\boldsymbol{\sigma}|\boldsymbol{\mu}, \tau\mathbf{V})$.

(c) Move to $\boldsymbol{\sigma}'$ given the current point $\boldsymbol{\sigma}$ with probability of move (Chib and Greenberg, 1995)

$$\min\left(\frac{\pi(\boldsymbol{\sigma}')\prod_{i=1}^{n}\prod_{t=1}^{T_i} f(y_{it}, s_{it}|\mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma}')}{\pi(\boldsymbol{\sigma})\prod_{i=1}^{n}\prod_{t=1}^{T_i} f(y_{it}, s_{it}|\mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma})}\,\frac{q(\boldsymbol{\sigma}|\boldsymbol{\mu}, \tau\mathbf{V})}{q(\boldsymbol{\sigma}'|\boldsymbol{\mu}, \tau\mathbf{V})}, 1\right),$$

otherwise stay at $\boldsymbol{\sigma}$. In the latter expression, $f(y_{it}, s_{it} = k|\mathbf{b}_i, \lambda_{it}, \boldsymbol{\beta}, \boldsymbol{\Sigma})$, $k = 0, 1$, is specified in (5) and (6).

(d) $\{\mathbf{z}_{it}\}$ from $\mathbf{z}_{it}|y_{it}, s_i, \boldsymbol{\beta}, \{\mathbf{b}_i\}, \boldsymbol{\Sigma}$, drawing each component of $\mathbf{z}_{it}$ independently for $i \leqslant n$. Following Albert and Chib (1993), if $s_{it}=1$, first sample s_{it}^* from the distribution $s_{it}^*|y_{it}, s_{it}, z_{it0}, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$, a normal distribution truncated to the interval $(0, \infty)$ and then given this draw of s_{it}^*, sample z_{it0} from the distribution $z_{it0}|y_{it}, s_{it}, s_{it}^*, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$, a normal distribution without any restriction on its support. If $s_{it} = 0$, modify the above so that s_{it}^* is sampled from the distribution $s_{it}^*|y_{it}, s_{it}, z_{it0}, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$, but now truncated to the interval $(-\infty, 0)$. Then given this draw sample z_{it1} from the distribution $z_{it1}|y_{it}, s_{it}, s_{it}^*, \mathbf{b}_i, \boldsymbol{\beta}, \boldsymbol{\Sigma}$.

2. Sample $(\boldsymbol{\beta}, \{\mathbf{b}_i\})$ from the distribution $(\boldsymbol{\beta}, \{\mathbf{b}_i\})|\mathbf{y}, \mathbf{s}, \{\mathbf{z}_{it}\}, \{\lambda_{it}\}, \boldsymbol{\gamma}, \mathbf{D}, \boldsymbol{\Sigma}$ by drawing

(a) $\boldsymbol{\beta}$ from $\boldsymbol{\beta}|\{\mathbf{z}_{it}\}, \{\lambda_{it}\}, \boldsymbol{\gamma}, \mathbf{D}, \boldsymbol{\Sigma}$, a Gaussian distribution with mean

$$\hat{\beta} = B\left(\boldsymbol{\beta}_0\mathbf{B}_0^{-1} + \sum_{i=1}^{n}\mathbf{X}_i'\boldsymbol{\Omega}_i^{-1}(\mathbf{z}_i - \mathbf{W}_i\mathbf{V}_i\boldsymbol{\gamma})\right)$$

and variance

$$\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i=1}^{n}\mathbf{X}_i\boldsymbol{\Omega}_i^{-1}\mathbf{X}_i\right)^{-1},$$

where $\boldsymbol{\Omega}_i = \mathbf{W}_i\mathbf{D}\mathbf{W}_i' + \boldsymbol{\Sigma}_i$ and $(\mathbf{X}_i, \mathbf{W}_i, \mathbf{z}_i)$ are defined in the discussion surrounding (7).

(b) $\{\mathbf{b}_i\}$ from $\mathbf{b}_i|\{\mathbf{z}_{it}\}, \{\lambda_{it}\}, \boldsymbol{\beta}, \boldsymbol{\gamma}, \mathbf{D}$, a Gaussian distribution with mean

$$\hat{\mathbf{b}}_i = \mathbf{C}_i\left(\mathbf{D}^{-1}\mathbf{V}_i\boldsymbol{\gamma} + \boldsymbol{\Sigma}^{-1}\sum_{t=1}^{T_i}\lambda_{it}(\mathbf{z}_{it} - \mathbf{X}_{it}\boldsymbol{\beta})\right)$$

and variance

$$\mathbf{C}_i = (\mathbf{D}^{-1} + \boldsymbol{\Sigma}^{-1}\lambda_i^*)^{-1},$$

where $\lambda_i^* = \boldsymbol{\Sigma}_{t=1}^{T_i}\lambda_{it}$.

3. Sample $\boldsymbol{\gamma}$ from $\boldsymbol{\gamma}|\{\mathbf{b}_i\}, \mathbf{D}$, a Gaussian distribution with mean $\hat{\boldsymbol{\gamma}} = \mathbf{C}(\mathbf{C}_0^{-1}\boldsymbol{\gamma}_0 + \sum_{i=1}^n \mathbf{V}_i' \mathbf{D}^{-1}\mathbf{b}_i)$ and variance $\mathbf{C} = (\mathbf{C}_0^{-1} + \sum_{i=1}^n \mathbf{V}_i'\mathbf{D}^{-1}\mathbf{V}_i)^{-1}$.
4. Sample $\mathbf{D}^{-1}$ from $\mathbf{D}^{-1}|\{\mathbf{b}_i\}, \boldsymbol{\gamma}$ where

$$\mathbf{D}^{-1}|\{\mathbf{b}_i\}, \boldsymbol{\gamma} \sim \text{Wishart}\left(\rho_0 + n, \left(\mathbf{R}_0^{-1} + \sum_{i=1}^{n}(\mathbf{b}_i - \mathbf{V}_i\boldsymbol{\gamma})(\mathbf{b}_i - \mathbf{V}_i\boldsymbol{\gamma})'\right)^{-1}\right).$$

5. Sample $\{\lambda_{it}\}$ by drawing λ_{it} from the distribution $\lambda_{it}|\mathbf{z}_{it}, \mathbf{b}_i, \boldsymbol{\lambda}^{(it)}, G_0, \boldsymbol{\beta}, \boldsymbol{\Sigma}$ where $\boldsymbol{\lambda}^{(it)}$ denotes the set of $\{\lambda_{it}\}$ excluding λ_{it}. Let $\phi^{(it)} = (\phi_1^{(it)}, \ldots, \phi_{n_{it}}^{(it)})$ denote the set of n_{it} unique values in the collection $\boldsymbol{\lambda}^{(it)}$ and let $m_j^{(it)}$ denotes the number of λ's in $\boldsymbol{\lambda}^{(it)}$ that take the value $\phi_j^{(it)}$. Then sample λ_{it} from the continuous-discrete distribution

$$\lambda_{it}|\mathbf{z}_{it}, \mathbf{b}_i, \boldsymbol{\lambda}^{(it)}, G_0, \boldsymbol{\beta}, \boldsymbol{\Sigma} \sim q_{it0}\pi_0(\lambda_{it}|\mathbf{z}_{it}, \boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma}) + \sum_{j=1}^{n_{it}} q_{itj}\delta(\phi_j^{(it)}),$$

where

$$\pi_0(\lambda_{it}|\mathbf{z}_{it}, \boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma}) = f_G\left(\lambda_{it}|, \frac{v+3}{2}, \frac{v + (\mathbf{z}_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - \mathbf{b}_i)'\boldsymbol{\Sigma}^{-1}(\mathbf{z}_{it} - \mathbf{X}_{it}\boldsymbol{\beta} - \mathbf{b}_i)}{2}\right)$$

is the gamma density and the weights are given by

$$\begin{aligned} q_{it0} &\propto \alpha \int f(\mathbf{z}_{it}|\boldsymbol{\beta}, \mathbf{b}_i, \boldsymbol{\Sigma})\, dG_0(\lambda_{it}) \\ &\propto \alpha f_T(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{b}_i, \boldsymbol{\Sigma}, v) \end{aligned}$$

and

$$q_{itj} \propto m_j^{(it)} f_N(\mathbf{z}_{it}|\mathbf{X}_{it}\boldsymbol{\beta} + \mathbf{b}_i, \boldsymbol{\Sigma}/\phi_j^{(it)}), \quad j = 1, \ldots, p_{it},$$

where f_T and f_N denote the multivariate-t and multivariate normal density functions, respectively.
6. Repeat Steps 1–5 using the most recent values of the conditioning variables.

References

Albert, J., Chib, S., 1993. Bayesian analysis of binary and polychotomous response data. Journal of the American Statistical Association 88, 669–679.

Albert, J., Chib, S., 1995. Bayesian residual analysis for binary response regression models. Biometrika 82, 747–759.

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91, 444–472.

Antoniak, C.E., 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. Annals of Statistics 2, 1152–1174.

Basu, S., Chib, S., 2001. Marginal likelihood and Bayes factors for Dirichlet process based mixture models. Manuscript, Washington University, St. Louis.

Bjorklund, A., Moffitt, R., 1987. The estimation of wage gains and welfare gains in self-selection models. Review of Economics and Statistics 69, 42–49.

Budd, J., Na, I., 2000. The union membership wage premium for employees covered by collective bargaining agreements. Journal of Labor Economics 18, 783–807.

Chib, S., 1992. Bayes inference in the Tobit censored regression model. Journal of Econometrics 51, 79–99.

Chib, S., 2001. Markov chain Monte Carlo methods: computation and inference. In: Heckman, J.J., Leamer, E. (Eds.), Handbook of Econometrics, Vol. 6. North-Holland, Amsterdam, pp. 3569–3649.

Chib, S., Carlin, B., 1999. On MCMC sampling in hierarchical longitudinal models. Statistics and Computing 9, 17–26.

Chib, S., Greenberg, E., 1995. Understanding the Metropolis–Hastings algorithm. American Statistician 49, 327–335.

Chib, S., Greenberg, E., 1996. Markov chain Monte Carlo simulation methods in econometrics. Econometric Theory 12, 409–431.

Chib, S., Greenberg, E., 1998. Analysis of multivariate probit models. Biometrika 85, 347–361.

Chib, S., Hamilton, B., 2000. Bayesian analysis of cross section and clustered data treatment models. Journal of Econometrics 97, 25–50.

Efron, B., Feldman, D., 1991. Compliance as an explanatory variable in clinical trials. Journal of the American Statistical Association 86, 9–26.

Escobar, M.D., West, M., 1995. Bayesian density estimation and inference using mixtures. Journal of the American Statistical Association 90, 577–588.

Fergusson, T.S., 1973. A Bayesian analysis of some nonparametric problems. Annals of Statistics 1, 209–230.

Freeman, R.B., 1984. Longitudinal analyses of the effects of trade unions. Journal of Labor Economics 2, 1–26.

Geweke, J., Keane, M., 2000. An empirical analysis of earnings dynamics among men in the PSID: 1968–1989 Journal of Econometrics 96, 293–356.

Heckman, J., Robb, R., 1985. Alternative methods for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, Cambridge.

Heckman, J., Ichimura, H., Todd, P., 1998. Matching as an econometric evaluation estimator. Review of Economic Studies 65, 261–294.

Hirano, K., 1999. Semiparametric Bayesian models for dynamic earnings data. Working paper, University of California, Los Angeles.

Hirano, K., Imbens, G., Ridder, G., 2000. Efficient estimation of average treatment effects using the estimated propensity score. National Bureau of Economic Research Technical Working Paper 251.

Jakubson, G., 1991. Estimation and testing of the union wage effect using panel data. Review of Economic Studies 58, 971–991.

Kyriazidou, E., 1997. Estimation of a panel data sample selection model. Econometrica 65, 1335–1364.

Lee, L.F., 1978. Unionism and wage rates: a simultaneous equation model with qualitative and limited dependent variables. International Economic Review 19, 415–433.

Lewis, H.G., 1986. Union Relative Wage Effects: A Survey. University of Chicago Press, Chicago.

MacEachern, S.N., Muller, P., 1998. Estimating mixture of Dirichlet process models. Journal of Computational and Graphical Statistics 7, 223–238.

Permutt, T., Hebel, J., 1989. Simultaneous-equation estimation in a clinical trial of the effect of smoking on birth weight. Biometrics 45, 619–622.

Ridder, G., 1992. Attrition in multi-wave panel data. In: Hartog, J., Ridder, G., Theeuwes, J. (Eds.), Panel Data and Labor Market Studies. Elsevier, North-Holland, Amsterdam.

Robinson, C., 1989. The joint determination of union status and union wage effects: some tests of alternative models. Journal of Political Economy 97, 639–667.

Rosenbaum, P., Rubin, D., 1983. The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Roy, A.D., 1951. Some thoughts on the distribution of earnings. Oxford Economic Papers 3, 135–146.

Rubin, D., 1974. Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66, 688–701.

Rubin, D., 1978. Bayesian inference for causal effects. The Annals of Statistics 6, 34–58.

Tanner, M.A., Wong, W.H., 1987. The calculation of posterior distributions by data augmentation. Journal of the American Statistical Association 82, 528–549.

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

Tiwari, R.C., Jammalamadaka, S.R., Chib, S., 1988. Bayes prediction density and regression estimation: a semi parametric approach. Empirical Economics 13, 209–222.

Vella, F., Verbeek, M., 1998. Whose wages do unions raise? A dynamic model of unionism and wage rate determination for young men. Journal of Applied Econometrics 13, 163–183.

Vella, F., Verbeek, M., 1999. Two-step estimation of panel data models with censored endogenous variables and selection bias. Journal of Econometrics 90, 239–263.

Wooldridge, J., 1995. Selection corrections for panel data models under conditional mean independence assumptions. Journal of Econometrics 68, 115–132.