ELSEVIER

Journal of Econometrics 97 (2000) 25–50

JOURNAL OF Econometrics

www.elsevier.nl/locate/econbase

Bayesian analysis of cross-section and clustered data treatment models

Siddhartha Chib*, Barton H. Hamilton

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

Received 1 May 1998; received in revised form 1 July 1998

Abstract

This paper is concerned with the problem of determining the effect of a categorical treatment variable on a response given that the treatment is non-randomly assigned and the response (on any given subject) is observed for one setting of the treatment. We consider classes of models that are designed for such problems. These models are subjected to a fully Bayesian analysis based on Markov chain Monte Carlo methods. The analysis of the treatment effect is then based on, amongst other things, the posterior distribution of the potential outcomes (counter-factuals) at the subject level, which is obtained as a by-product of the MCMC simulation procedure. The analysis is extended to models with categorical treatments and binary and clustered outcomes. The problem of model comparisons is also considered. Different aspects of the methodology are illustrated through two data examples.

JEL classification: C1; C4

Keywords: Causal inference; Categorical treatments; Finite mixture distribution; Gibbs sampling; Marginal likelihood; Markov chain Monte Carlo; Non-experimental data; Potential outcomes; Randomly assigned covariate; Sample selection; Treatment effect

* Corresponding author.
E-mail address: chib@olin.wustl.edu (S. Chib).

0304-4076/00/$ - see front matter
PII: S0304-4076(99)00065-2

1. Introduction

This paper is concerned with the problem of causal inference in models with non-randomly assigned treatments. Models with this feature have been widely analyzed in the econometrics and statistics literatures, often with different nomenclatures to reflect the setting of the problem (Copas and Li, 1997; Efron and Feldman, 1991; Imbens and Rubin, 1997; Heckman, 1978; Heckman and Robb, 1985; Holland, 1986; Lee, 1979; Maddala, 1983; Rubin, 1974,1978). For example, models of sample selection and selectivity arise frequently in economics while the problem of compliance (and dropouts) is crucial in connection with interpreting the results of a clinical trial. Both settings deal with the non-random assignment of 'treatments'. In the selection problem, the receipt of the treatment and the observed outcome are correlated due to some unmodeled subject-specific factors. As a result, the true treatment effect is confounded, unless some attempt is made to mitigate the selection problem. In the biostatistics context, the true effect of the drug on the outcome, relative to the placebo, is generally confounded if unobserved factors that influence compliance, or dropouts, are correlated with the outcome.

One purpose of this paper is to offer a flexible Bayesian analysis of these problems in settings that are more general than the ones that have been considered in the literature. For example, Wooldridge (1995) and Kyriazidou (1997) consider panel data models in which selection is binary and the single outcome is continuous. By contrast, in one of our problems, the treatment variable is ordinal and the outcomes are binary and clustered. Another purpose of the paper is to extract subject-specific treatment effects, as opposed to the mean (population averaged) treatment effect. The importance of treatment effect heterogeneity has been emphasized by Heckman (1997). That this can be done is a consequence of both our estimation framework (which relies on Markov chain Monte Carlo methods) and on our use of models that explicitly involve potential outcomes (counter-factuals). As far as we are aware, the computation of subject-specific posterior treatment distributions is new. Finally, the paper develops a procedure for comparing alternative treatment models. This comparison is done on the basis of marginal likelihoods that are computed by the method of Chib (1995).

The rest of the paper is organized as follows. In Section 2 we present the prior–posterior analysis of models in which the treatment is binary and the outcomes are continuous. Section 3 extends the basic methodology to random effects clustered data models with ordinal treatments and binary outcomes. Section 4 discusses the computation of the subject-level treatment effect distributions and shows how the marginal likelihood of competing models can be obtained from the simulation output. Sections 5 and 6 are concerned with the details of the data analysis while Section 7 concludes.

2. Binary treatment and continuous outcomes

To fix ideas, we begin with the textbook problem consisting of a binary treatment $s \in \{0,1\}$ and a Gaussian outcome $y \in \Re$. On the ith subject in the sample ($i \leqslant n$), we observe the data $(x_i, w_i, s_i = k, y_i)$, where x_i and w_i are covariates, k is either zero or one and y_i denotes the response given that s_i is equal to k.

Now let z_{i0} and z_{i1} denote the *potential* outcomes for each level of the covariate s_i and suppose that

$$\begin{pmatrix} s_i^* \\ z_{i0} \\ z_{i1} \end{pmatrix} \sim \mathcal{N}_3 \left(\begin{pmatrix} w_i'\gamma \\ x_{i0}'\beta_0 \\ x_{i1}'\beta_1 \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} \\ \sigma_{12} & \sigma_{22} & \sigma_{23} \\ \sigma_{13} & \sigma_{23} & \sigma_{33} \end{pmatrix} \right) \tag{1}$$

or compactly as

$$z_i \sim^{\text{ind}} \mathcal{N}_3(X_i\beta, \Sigma), \quad i = 1,2,\ldots,n,$$

where w_i: $p \times 1$, x_{i0}: $k_0 \times 1$ and x_{i1}: $k_1 \times 1$ are subsets of the covariate vector x_i, $X_i = \text{diag}(w_i', x_{i0}', x_{i1}')$, ind denotes independence, $\mathcal{N}_k$ the k-variate normal distribution, and $\beta = (\gamma, \beta_0, \beta_1)$: $k \times 1$ ($k = p + k_0 + k_1$). The observed treatment and observed outcomes are given by

$$s_i = I[s_i^* > 0]$$

and

$$y_i = \begin{cases} z_{i0} & \text{if} \quad s_i = 0, \\ z_{i1} & \text{if} \quad s_i = 1, \end{cases}$$

respectively. Thus, only one of the outcomes z_{i0} and z_{i1} is observed, depending on the value taken by s_i. In this model, w_i contains at least one covariate (instrument) r_i, not present in x_{i0} or x_{i1}, that is genuinely randomly assigned. The presence of a covariate r_i is crucial to extracting the treatment effect, as is well known. The parameters of this model are β and $\sigma = (\sigma_{12}, \sigma_{22}, \sigma_{13}, \sigma_{33})$, with σ_{11} set to one because the scale of s_i^* is not determinate, and σ_{23} to zero because σ_{23} does not appear in the likelihood function (see Koop and Poirier, 1997 on the implications of not imposing the latter restriction).

2.1. Extensions of basic model

This textbook model, which has been widely studied, is known to be sensitive to the Gaussian assumption. A substantial body of work directed at overcoming this problem has now appeared, based almost entirely on a semiparametric viewpoint (Newey et al., 1990; Ahn and Powell, 1993; Kyriazidou, 1997). This literature, unfortunately, has not had a significant impact on the empirical fitting of these models.

In this paper we proceed in a different direction by considering classes of flexible parametric models that relax the Gaussian assumption but maintain tractability. Most importantly, this framework, as we show below, can be extended to panel data and ordinal treatment problems for which, at this time, no semiparametric methods are available.

One simple elaboration is to let $z_i = (s_i^*, z_{i0}, z_{i1})$ follow a multivariate-t distribution. Let $\{\lambda_i,\ i \leqslant n\}$ be independently distributed random variables from a gamma $\mathcal{G}(\nu/2,\ \nu/2)$ distribution and let

$$z_i|\beta,\ \theta,\ \lambda_i \sim^{\text{ind}} \mathcal{N}_3(X_i\beta,\ \lambda_i^{-1}\Sigma). \tag{2}$$

Then, unconditionally on λ_i, z_i follows a multivariate-t distribution with ν degrees of freedom and density

$$f(z_i|\beta,\ \theta) \propto |\Sigma|^{-1/2}\left(1 + \frac{1}{\nu}(z_i - X_i\beta)'\Sigma^{-1}(z_i - X_i\beta)\right)^{-(3+\nu)/2}.$$

The observed treatments and the responses are given as before.

Another simple extension is to let $z_i = (s_i^*, z_{i0}, z_{i1})$ follow a mixture of multivariate Gaussian distributions. Let m denote the number of components in the mixture and let v_i be a discrete random variable (taking values in $\{1, 2, \ldots, m\}$) that represents the component from which the ith observation is drawn. In particular, $v_i = j$ specifies that the ith observation is drawn from the jth component population. Then, the mixture model is given by

$$z_i|\beta,\ \sigma,\ v_i = j \sim \mathcal{N}_3(X_i\beta^j,\ \Sigma^j), \tag{3}$$

where we have let each component possess its own regression vector $\beta^j = (\gamma^j,\ \beta_0^j,\ \beta_1^j)$ and covariance matrix Σ^j. The parameters β and σ now denote the complete set of $\{\beta^j\}$ and $\{\sigma^j\}$, respectively. Under the assumption that $\Pr(v_i = j) = q_j$, it follows that the distribution of the z_i is given by the mixture of Gaussian distributions

$$(s_i^*, z_{i0}, z_{i1})|\beta,\ \sigma \sim \sum_{j=1}^{m} q_j \mathcal{N}_3(X_i\beta^j,\ \Sigma^j) \tag{4}$$

and the model is completed by the specification of the observed data, as before.

2.2. Prior distributions for parameters

The prior on β is multivariate Gaussian and is generically denoted as $\mathcal{N}_k(\beta_0, B_0)$. In the case of the mixture model in (4), the β^j's are modeled as exchangeable and assumed to be drawn from a common $\mathcal{N}_k(\beta_0, B_0)$ population distribution. For the covariance parameter σ, we need to ensure that the distribution is truncated to the region $S \subset \Re^4$ that leads to a positive-definite matrix Σ. Following Chib and Greenberg (1998), we let the prior distribution

be truncated normal $\sigma \propto \mathcal{N}_4(g_0, G_0)I_S(\sigma)$ where g_0 and G_0 denote the hyperparameters and I_S the indicator function taking the value one if σ is in S and the value zero otherwise. This prior is flexible and convenient and can be used to incorporate various prior beliefs about the variances and the covariances. For the mixture model, we assume that independently $\sigma^j \propto \mathcal{N}_4(g_0, G_0)I_S(\sigma)$ where σ^j denotes the free parameters of Σ^j ($j \leqslant m$).

In general, the hyperparameters (β_0, B_0, g_0, G_0) of these prior distributions must be assigned subjectively. In some cases, the hyperparameters may be based on the estimates from a prior (historical) training sample. The posterior mean of the parameters (under default priors) from the training sample data can be used to specify β_0 and g_0 and the associated posterior covariance matrix (inflated to account for the differences between the training sample and the current data) may be used to specify B_0 and G_0.

2.3. Prior–posterior analysis

The posterior distributions of the parameters (β, σ) in the models presented cannot be summarized by analytical means due to the complexity of the likelihood functions and the restrictions on the parameters. These posterior distributions are, however, amenable to analysis by simulation-based methods, in particular those based on Markov chain Monte Carlo (MCMC) methods. Within this framework, the general idea is to base inferences on a (correlated) sample of draws from the posterior distribution where the sample is obtained by simulating a suitably constructed (high-dimensional discrete-time continuous state space) Markov chain whose invariant distribution is the desired posterior distribution (see Chib and Greenberg, 1995; Tanner and Wong, 1987; Tierney, 1994 for more details of these methods). As we show below, the simulation steps are aided by an augmentation of the parameter space to include the latent potential outcomes (following Tanner and Wong, 1987; Chib, 1992 and Albert and Chib, 1993). To improve the behavior of the Markov chain, we base the simulation steps on a reduced blocking scheme in which the free elements of the covariance structure are simulated from a conditional distribution that is marginalized over the potential outcomes.

2.3.1. Gaussian model

Consider first the textbook model, where the treatment is binary and the potential outcomes are Gaussian. The likelihood function of this model is

$$
\begin{aligned}
f(y, s|\beta, \sigma) &\propto f(y|\beta, \Sigma)\Pr(s|y, \beta, \Sigma) \\
&\propto \prod_{i:s_i=0} f_{\mathrm{N}}(z_{i0}|x_{i0}'\beta_0, \sigma_{22}^2)\Phi\left(\frac{-w_i'\gamma - (\sigma_{12}/\sigma_{22})(z_{i0} - x_{i0}'\beta_0)}{(1-\sigma_{12}^2/\sigma_{22})^{1/2}}\right) \\
&\quad \times \prod_{i:s_i=1} f_{\mathrm{N}}(z_{i1}|x_{i1}'\beta_1, \sigma_{33}^2)\Phi\left(\frac{w_i'\gamma + (\sigma_{13}/\sigma_{33})(z_{i1} - x_{i1}'\beta_1)}{(1-\sigma_{13}^2/\sigma_{33})^{1/2}}\right),
\end{aligned} \tag{5}
$$

where $\beta = (\gamma, \beta_0, \beta_1) \in \Re^k$, $\sigma = (\sigma_{12}, \sigma_{13}, \sigma_{22}, \sigma_{33}) \in S$, f_N denotes the normal density function and Φ is the cdf of the standard normal distribution. The posterior density of (β, σ) is given by

$$\pi(\beta, \sigma|y, s) \propto \pi(\beta)\pi(\sigma)f(y, s|\beta, \sigma), \tag{6}$$

where the prior distributions of β and σ are $\mathcal{N}_k(\beta_0, B_0)$ and $\mathcal{N}_4(g_0, G_0)I_S(\sigma)$, respectively. Under mild assumptions that are similar to those in Chib (1992) for the Tobit model, one can show that the posterior density is proper. The arguments, which are straightforward, rely on the fact that the cdf terms are uniformly bounded for all admissible values of the parameters.

We now describe a strategy for sampling the posterior density in (6). If we let z_i^*: 2×1 denote the unobserved components of $z_i = (s_i^*, z_{i0}, z_{i1})$, then our Markov chain Monte Carlo simulations are conducted on the data augmented posterior density $\pi(z_1^*, \ldots, z_n^*, \beta, \sigma|y, s)$. To improve the efficiency of the algorithm, we use a reduced (as opposed to full) blocking structure. We summarize the steps in Algorithm 1 and defer the details to the appendix.

Algorithm 1

1. Initialize β
2. Sample σ and $\{z_i\}$ from $\pi(\sigma, \{z_i\}|y, s, \beta)$ by sampling
   (a) σ from $\pi(\sigma|y, s, \beta)$ using the Metropolis–Hastings algorithm and
   (b) $\{z_i\}$ from $\pi(\{z_i^*\}|y, s, \beta, \Sigma)$
3. Sample β from $\pi(\beta|y, s, \{z_i^*\}, \Sigma) = \pi(\beta|\{z_i^*\}, \Sigma)$.
4. Repeat Steps 2–3 using the most recent values of the conditioning variables.

The reduced blocking step consists of the joint sampling of σ and $\{z_i^*\}$ from the conditional posterior density $\pi(\sigma, \{z_i^*\}|y, s, \beta)$. This tactic tends to reduce the serial correlation in the MCMC output, hence increasing the simulation accuracy of the estimates based on the posterior sample.

2.3.2. Student-t model

Now consider the case of the t-model in which the treatment intensity s_i^* and the potential outcomes follow the multivariate-t distribution. To deal with this situation, one can adopt the strategy discussed by Albert and Chib (1993) and augment the parameter space by the Gamma mixing variables λ_i that appear in (2). Then, conditionally on the values of $\{\lambda_i\}$, the distribution of z_i is Gaussian with covariance matrix $\lambda_i^{-1}\Sigma$ and the parameters $(\beta, \sigma, \{z_i^*\})$ can be simulated as in Algorithm 1. The MCMC scheme for this model is completed by sampling λ_i, given $(\beta, \sigma, \{z_i^*\})$ and the data, from the distribution

$$\mathcal{G}\left(\frac{\nu + 3}{2}, \frac{\nu + (z_i - X_i\beta)'\Sigma^{-1}(z_i - X_i\beta)}{2}\right),$$

independently for $i = 1, 2, \ldots, n$.

2.3.3. Mixture model

The basic MCMC scheme outlined for the Gaussian model can also be adapted to deal with the case in which the potential outcomes follow a finite mixture of Gaussian distributions. Following the approach of Diebolt and Robert (1994) for mixture models, we utilize the representation of the model in (3) and augment the parameter space by the discrete-valued component indicator variables $\{v_i\}$. Conditioned on the values of $\{v_i\}$, the data separate into m blocks, with each block consisting of the observations that are ascribed to component j, $j = 1, \ldots, m$. Given these blocks of observations, the parameters of the jth component, namely $(\beta^j, \sigma^j, \{z_i^{*j}\})$, can be updated according to the MCMC scheme described for the Gaussian model. Next, given the updated values of the parameters, a new value of v_i $(i = 1, 2, \ldots, n)$ is simulated from the discrete mass distribution

$$\Pr(v_i = j|y_i, s_i, \beta, \sigma, z_i^*) \propto q_j|\Sigma^j|^{-1/2} \exp\{-0.5(z_i - X_i\beta^j)'(\Sigma^j)^{-1} \times (z_i - X_i\beta^j)\} \quad j \leqslant m.$$

In this MCMC scheme we do not address the choice of m or the local non-identifiability of the mixture distribution to relabeling of the component indicators. Our view of the former issue is that the choice of m is a problem of model comparison that is best addressed by the methods in Section 4. As for the latter, we view the mixture model as providing a semi-parametric modeling of the potential outcomes and the relabeling problem is of concern if more components than are necessary are used. Besides, since we focus on the posterior distributions of the potential outcomes marginalized over the parameters, the local non-identifiability of the mixture labels can be ignored.

3. Clustered data with selection

Consider now treatment models that are appropriate for unbalanced clustered data problems. Clustered data may arise from a panel study or, as in our example in Section 6, from a specified grouping of subjects in a cross-sectional problem. In such settings, it is likely that subjects in a given cluster are likely to receive the same treatment and to also have some similarities in the outcomes. Thus, it becomes necessary to model the cluster-specific effects in conjunction with the selection effect. To capture the features of the data that we analyze below, we let the treatment variable be ordinal and the outcome variable be binary. Previous panel data selection models, e.g., Wooldridge (1995), Kyriazidou (1997) and Heckman et al. (1998), deal exclusively with the binary selection, continuous outcome case. In addition, except for the last of these, the papers do not consider potential outcomes.

Let l denote the lth cluster, $l = 1, \ldots, m$, which consists of n_l subjects. In our example below, a cluster is defined by a hospital, and n_l denotes the number of sample patients treated in the lth hospital. Suppose that the treatment variable for the ith subject in the lth cluster s_{li} is ordinal taking possible values $\{0,1,\ldots,J\}$. For each value of s_{li}, let $d_{lik} \in \{0,1\}$ be a binary *potential* response random variable. The response d_{lik} is observed if $s_{li} = k$ and is not observed otherwise.

The model of interest is now given by

$$\begin{pmatrix} s_{li}^* \\ z_{li0} \\ \vdots \\ z_{liJ} \end{pmatrix} \sim \mathcal{N}_{J+2} \left(\begin{pmatrix} w_{li}'\gamma + a_l \\ x_{li0}'\beta_0 + b_{l0} \\ \vdots \\ x_{liJ}'\beta_J + b_{lJ} \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} & \cdots & \sigma_{1J+2} \\ \sigma_{12} & 1 & 0 & 0 \\ \vdots & 0 & \ddots & \vdots \\ \sigma_{1J+2} & 0 & \cdots & 1 \end{pmatrix} \right) \tag{7}$$

or compactly as

$$z_{li}|b_l \sim^{\text{ind}} \mathcal{N}_{J+2}(X_{li}\beta + b_l, \Sigma),$$

where $z_{li} = (s_{li}^*, z_{li0}, \ldots, z_{liJ})$, $X_{li} = \text{diag}(w_{li}', x_{li0}', \ldots, x_{liJ}')$, $\beta = (\gamma', \beta_0', \ldots, \beta_J)'$ and $b_l = (a_l, b_{l0}, \ldots, b_{lJ})$: $J + 2 \times 1$ is the vector of cluster-specific random effects with distribution

$$b_l \sim^{\text{ind}} N_{J+2}(0, D)$$

with D a $(J + 2) \times (J + 2)$ full positive-definite matrix. Note that we have let each response have its own random effect. The observed treatments are generated according to

$$s_{li} = \begin{cases} 0 & \text{if} \quad s_{li}^* < 0 \\ j & \text{if} \quad \xi_{j-1} \leqslant s_{li}^* < \xi_j, \quad j = 2, \ldots, J-1, \\ J & \text{if} \quad \xi_{J-1} \leqslant s_{li}^*, \end{cases} \tag{8}$$

which implies that conditionally on the random effects, the observed treatment for the ith subject in the lth cluster is given by the ordinal probit model with probabilities $\Pr(s_{li} \leqslant j) = \Pr(s_{li}^* < \xi_j) = \Phi(\xi_j - w_{li}'\gamma - a_l)$. In addition, one observes the outcome

$$y_{li} = d_{lik} \quad \text{if} \quad s_{li} = k$$

with

$$d_{lik} = \begin{cases} 1 & \text{if} \quad z_{lik} > 0, \\ 0 & \text{if} \quad z_{lik} \leqslant 0 \end{cases} \tag{9}$$

and the outcome d_{lik} is not observed otherwise.

If we let $\theta = (\beta, \xi, \Sigma, D)$ where $\xi = (\xi_1, \ldots, \xi_{J-1})$ is a $J-1$ vector of cut-points and D is the matrix of variances and covariances of the random effects, then the likelihood contribution of the observations in the lth cluster is given by

$$f(y_l, s_l|\theta) = \int \left\{ \prod_{i=1}^{n_{l_i}} f(y_{li}, s_{li}|b_l, \theta) \right\} \pi(b_l|D^*)\, \mathrm{d}b_l. \tag{10}$$

This integral is complicated and, in general, is best computed by simulation methods. A particular approach based on importance sampling is described below.

We mention that the diagonal elements of the covariance matrix Σ are unity because (given the binary responses) the scale of z_{li} is not determinate. In addition, the covariance matrix of the latent potential outcomes $(z_{li0}, \ldots, z_{liJ})$ is the identity matrix because none of these parameters enter the data likelihood function. It should be noted that this restriction (in contrast with the cross-section case) is less material because of the correlation in potential outcomes that is induced by the random effects. The only unknown parameters in the Σ matrix are, therefore, in the first row which we denote by $\sigma = (\sigma_{12}, \ldots, \sigma_{1J+2})$ that again must lie in a region S that leads to a positive-definite covariance matrix. Note that we economize on notation by using the same symbols as in the previous section and the meaning of β, σ and S, for example, depends on the context.

3.1. Prior distributions

We assume that β is $\mathcal{N}_k(\beta_0, B_0)$ and that $\sigma \propto \mathcal{N}_{J+1}(g_0, G_0) I_S(\sigma)$. To deal with the ordered cut-points ξ, we adopt the approach of Albert and Chib (1997) and reparameterize the cut-points as

$$\alpha_1 = \log \xi_1, \quad \alpha_j = \log(\xi_j - \xi_{j-1}), \quad 2 \leqslant j \leqslant J-1, \tag{11}$$

with inverse map given by

$$\xi_j = \sum_{i=1}^{j} \exp(\alpha_i), \quad 1 \leqslant j \leqslant J-1.$$

We then assign α an unrestricted multivariate Gaussian prior distribution with mean α_0 and covariance matrix A_0. Finally, for the variance matrix D we let D^{-1} follow a Wishart distribution with parameters v_0 and R_0 (implying that the prior mean of D^{-1} is $v_0 R_0$). The hyperparameters of these prior distributions are assigned subjectively. In our application, however, the hyperparameters are found from a training sample.

3.2. Posterior sampling

Once again (due to the complexity of the likelihood function) it is necessary to augment the parameter space with the latent data to effectively sample the posterior distribution. We include both $\{z_i\}$ and $\{b_l\}$ in the sampler and adopt a reduced blocking scheme in which α and σ are sampled in one block conditioned on the cluster-specific random effects $b_l = (a_l, b_{l0}, \ldots, b_{lJ})$, but marginalized over the latent data $\{z_i\}$. In particular, the sampling of (α, σ) is from the density

$$\pi(\alpha, \sigma|y, s, \{b_l\}, \beta) \propto \pi(\alpha)\pi(\sigma)\prod_{l=1}^{m} f(s_l, y_l|\alpha, \beta, \{b_l\}, \Sigma)$$

$$\propto \pi(\alpha)\pi(\sigma)\prod_{l=1}^{m}\prod_{i=1}^{n_l} f(s_{li}, y_{li}|\alpha, \beta, b_l, \Sigma), \tag{12}$$

where $f(s_l, y_l|\alpha, \beta, \Sigma)$ is the likelihood contribution from the lth cluster and $f(s_{li}, y_{li}|\alpha, \beta, b_l, \Sigma)$ is the density of the ith observation in the lth cluster that is computed as follows. Suppose that the data on the ith subject in the lth cluster is $(s_{li}, y_{li}) = (k,1)$, which implies that s_{li}^* is between $\xi_{k-1} \leqslant s_{li}^* < \xi_k$ and that z_{lik} is positive, since all other potential outcomes can be integrated out and play no role in the probability calculation. Letting $\Phi_2(t_1, t_2; \rho)$ denote the cdf of the standard bivariate Gaussian distribution with mean zero, unit variances and correlation ρ, it follows that

$$f(k, 1|\alpha, \beta, b_l, \Sigma) = \Pr(\xi_{k-1} \leqslant s_{li}^* < \xi_k, 0 < z_{lik} < \infty)$$

$$= \Phi(\xi_k - w_{li}'\gamma) - \Phi(\xi_{k-1} - w_{li}'\gamma)$$

$$- \Phi_2(\xi_k - w_{li}'\gamma, -x_{lik}'\beta_k, \sigma_{1k+2})$$

$$+ \Phi_2(\xi_{k-1} - w_{li}'\gamma, -x_{lik}'\beta_k, \sigma_{1k+2}), \tag{13}$$

from standard properties of the bivariate normal distribution function. In doing this calculation one must be careful to ensure that $s_{li} = k$ is associated with the correct potential outcome z_{lik}. The posterior density in (12), which appears quite formidable, can be sampled effectively by a tuned Metropolis–Hastings step along the lines of Chib and Greenberg (1998), as discussed fully in the appendix.

Another important issue in the posterior computations for this model is the sampling of β and $\{b_i\}$. It has been emphasized by Chib and Carlin (1999) for general hierarchical models that, whenever possible, the fixed and random effects should be sampled in one block. Conditioned on the potential outcomes and the covariance parameters the clustered data model given above reduces to

a (linear) multivariate model for which the joint sampling of $(\beta, \{b_l\})$ can be conducted easily by the method of composition, by the sampling of β marginalized over $\{b_i\}$, followed by the sampling of $\{b_i\}$ given β.

The full MCMC algorithm (containing both reduced sampling steps) is summarized as follows. Details are furnished in the appendix.

Algorithm 2

1. Initialize $\{b_l\}$ and β
2. Sample $(\alpha, \sigma, \{z_{li}\})$ from $\alpha, \sigma, \{z_{li}\}|y, s, \{b_l\}, \beta$ by drawing
   (a) (α, σ) from $(\alpha, \sigma)|y, s, \{b_l\}, \beta$ using the Metropolis–Hastings algorithm and
   (b) $\{z_{li}\}$ from $\{z_{li}\}|s_l, \beta, \{b_l\}, \alpha, \Sigma$;
3. Sample $(\beta, \{b_l\})$ from $\beta, \{b_l\}|\{z_{li}\}, \Sigma$ by drawing
   (a) β from $\beta|\{z_{li}\}, \Sigma$ and
   (b) b_l from $b_l|\{z_{li}\}, \beta, D$ $(l \leqslant m)$;
4. Sample D from $D|\{b_l\}$
5. Repeat Steps 2–4 using the most recent values of the conditioning variables.

4. Posterior inferences

4.1. Inferring the treatment effect

One of the key issues in the prior posterior analysis is the question of inferring the treatment effect given one of the models we have just specified. See also Heckman and Robb (1985) and Angrist et al. (1996) for discussion of these matters.

Consider the case in which there is a binary treatment s_i, an instrumental variable that influences s_i but not the response, and two potential outcomes z_{i0} and z_{i1} that depend on the observed treatment. The causal effect of s on the response y for subject i is the difference $z_{i1} - z_{i0}$. If subject i receives the treatment $s_i = 0$, then z_{i0} is observed and the treatment effect is the quantity $z_{i1} - y_i$. On the other hand, when $s_i = 1$, z_{i1} is observed and the treatment effect is the quantity $y_i - z_{i0}$. Thus, with binary treatments, the treatment effect T_i is

$$T_i = \begin{cases} z_{i1} - y_i & \text{if} \quad s_i = 0, \\ y_i - z_{i0} & \text{if} \quad s_i = 1. \end{cases}$$

In either case, the treatment effect is a *random*, subject-specific quantity. From a Bayesian perspective, inference about the treatment effect requires the calculation of the posterior distribution of $\{z_{i1} - z_{i0}\}$ conditioned on the data

(y_i, X_i, w_i) and the treatment s_i but marginalized over the posterior distribution of parameters. This view, which is a consequence of our Bayesian approach, is a departure from that discussed by Angrist, Imbens and Rubin (1996) who summarize the treatment effect (conditioned on the parameters) by finding the mean treatment effect for different sub-populations defined by values of the outcomes and the (discrete) instrument.

We focus on the posterior distribution of the T_i although it should be noted that given T_i the treatment effects aggregated over various sub-populations (for example, the treated) can easily be constructed. The information in the subject-level posterior distributions of T_i can also be summarized in other ways. One can, for example, compute the posterior mean of z_{i1} and plot the distribution of these posterior means across the subjects in the sample. This can be compared with the corresponding distribution computed from the posterior means of $\{z_{i0}\}$. The important point is that the potential outcomes framework, in conjunction with the Bayesian paradigm, does not lead to a unique numerical summary of the treatment effect.

Another interesting set of questions arise in trying to calculate the treatment effect in the context of the ordinal treatment models and binary treatments. In these cases, if the treatment takes three levels (say), then there are three possible treatment effects for each subject, corresponding to the differences $z_{i1} - z_{i0}$, $z_{i2} - z_{i0}$ and $z_{i2} - z_{i1}$. One can define the treatment effects to be the posterior distribution of these subject-specific differences in potential outcomes. It should be noted that even if the response is binary as in the clustered data model, the modeling approach discussed above leads to a posterior distribution on continuous-valued potential outcomes. A cautionary remark is that in the latter case the distribution of these differences may be sensitive to the identifying (zero) covariance restriction on the lower part of the Σ matrix although note our earlier comment that this is mitigated by the correlation induced by the random effects. One could instead define the treatment effect in terms of the differences of binary potential outcomes but at the cost of a less interpretable quantity.

4.2. Computation of the marginal likelihood

In the fitting of selection models an important question relates to how alternative models can be compared. Such models may arise from the covariate specification in the treatment and potential outcomes distributions or from different distributional assumptions and modeling features (for example, the presence of clustering versus no clustering). To compare these models, one must compute the model marginal likelihood which is defined as the integral of the sampling density with respect to the prior density. This quantity can be computed from the MCMC output using a method developed by Chib (1995).

Let θ denote the parameters of a given model, with likelihood function $f(y, s|\theta)$ and prior density $\pi(\theta)$, where (y, s) is the available data. The parameters θ, the

likelihood and the prior are model dependent but that is not emphasized in the notation. Then, the marginal likelihood can be written as

$$m(y,s) = \frac{f(y,s|\theta)\pi(\theta)}{\pi(\theta|y,s)},$$

which follows from the formula for the posterior density of θ. The important point is that this expression is an identity in θ and may therefore be evaluated at any appropriately selected point θ^* (say). If θ^* denotes a high-density point and $\hat{\pi}(\theta^*|y,s)$ the estimate of the posterior ordinate at θ^*, then the marginal likelihood on the log scale is estimated as

$$\ln \hat{m}(y,s) = \ln f(y,s|\theta^*) + \ln \pi(\theta^*) - \hat{\pi}(\theta^*|y,s), \tag{14}$$

where the first two terms are typically available directly and the third is estimated from the MCMC output. We now very briefly explain how the first and third terms are determined for our various models.

There is little difficulty in finding the likelihood function $f(y,s|\theta^*)$ for the continuous outcome models. For example, in the textbook Gaussian model, the likelihood function is given by (5). For the clustered data model, the likelihood contribution of the lth cluster takes the form

$$\begin{aligned} f(y_l, s_l|\theta^*) &= \int \left\{ \prod_{i=1}^{n_{li}} f(y_{li}, s_{li}|b_l, \theta^*) \right\} \pi(b_l|D^*)\, \mathrm{d}b_l, \\ &\equiv \int g(b_l)\, \mathrm{d}b_l, \end{aligned}$$

where each term $f(y_{li}, s_{li}|b_l, \theta^*)$, by virtue of being conditioned on b_l, is found in the same way as described in the discussion surrounding (13). This multi-dimensional integral over $b_l = (a_l, b_{l0}, \ldots, b_{lJ})$ can be estimated by the method of importance sampling. Let $\hat{b}_l$ denote the mode of $\ln g(b_l)$ and V_{b_l} the inverse of minus the Hessian matrix at the mode. Then, if $h(b_l)$ denotes a multivariate-t density with mean $\hat{b}_l$, scale aV_{b_l} and v degrees of freedom ($a > 1$ and v are tuning factors), the importance sampling estimate of the likelihood contribution is given by $T^{-1}\sum_{t=1}^{T} g(b_l^t)/h(b_l^t)$, where $\{b_l^t\}$ are random draws from the importance density.

Next consider the estimation of the posterior ordinate $\pi(\theta^*|y,s)$. The main ideas can be illustrated in the context of the textbook model where $\theta = (\beta, \sigma)$. Decompose the posterior ordinate according to the marginal/conditional decomposition as

$$\pi(\sigma^*, \beta^*|y,s) = \pi(\sigma^*|y,s)\pi(\beta^*|y, s, \sigma^*)$$

and note that the first ordinate $\pi(\sigma^*|s, y)$ can be estimated by kernel smoothing using the draws $\{\sigma^{(g)}\}$ from the MCMC output, as in Chib and Greenberg (1998). The second ordinate by definition is

$$\pi(\beta^*|y, s, \Sigma^*) = \int \pi(\beta^*|\{z_i\}, \Sigma^*)\, \mathrm{d}\pi(\{z_i\}|y, s, \Sigma^*),$$

where the integrating measure is the posterior distribution of the latent data conditioned on Σ^* and the integrand $\pi(\beta^*|\{z_i\}, \Sigma^*)$ is the ordinate of the Gaussian density defined in Section 3.1. Following Chib (1995), to estimate this integral one fixes the value of σ at σ^* (equivalently Σ at Σ^*) and continues the MCMC iterations with the reduced set of full conditional distributions $\beta|\{z_i\}, \Sigma^*$ and $\{z_i\}|y, s, \beta, \Sigma^*$. The draws from this run on $\{z_i\}$ are then used to average the Gaussian ordinate $\pi(\beta^*|\{z_i\}, \Sigma^*)$.

The calculation of the posterior ordinate in the other models proceeds in the same fashion with some straightforward modifications.

5. Example 1: Hospice data

This example is concerned with the fitting of the models in Section 2 and their comparison through marginal likelihoods and Bayes factors, and the estimation of subject-specific treatment effects.

5.1. Data construction

A random sample of data for 1987 was collected on 568 United States hospices (institutions that provide care to the terminally ill patients) to determine the effect of certification under the Medicare Hospice Benefit (MHB) program on the number of patients served by the hospice (the outcome). The treatment variable s_i is 1 if the hospice is certified and 0 otherwise and z_{i0} is the natural logarithm of the number of patients served by the hospice when $s_i = 0$. The other potential outcome z_{i1} corresponds to $s_i = 1$.

Assume that $x_{i0} = x_{i1}$, where x_{i0} consists of (1) the number of years the hospice has operated; (2) per-capita income in the county in which the hospice is located; (3) average length of stay at the hospice and (4) percent of female patients. Also we let the marginal distribution of s_i be

$$\Pr(s_i = 1) = \Phi(w_i\gamma),$$

where w_i consists of x_{i0}, plus the MHB reimbursement rate, the Medicare Home Health reimbursement rate, the average salary for hospital employees in that

county, and the Health Care Finance Administration's (HCFA) labor cost index. This specification is based on Hamilton (1993).

5.2. Model selection and posterior analysis

A number of the different selection models discussed in Section 2 are fit to these data using Algorithm 1 and its variants — these are the Gaussian model, the Student-t model with ν equal to 4, 8, 16, 32, and 64, and mixture models with two and three components. Each model was fit using 10,000 MCMC iterations with an initial burn-in of 1000 iterations.

The MCMC output from the fitting is quite well behaved. In Fig. 1 we report the posterior histogram and time-series autocorrelation plots based on the output of σ_{13} and σ_{33} from the $\nu = 8$ model. For contrast we also report the corresponding acf plots with full blocking (i.e., σ is sampled conditioned on $\{z_i\}$). The plots show that the serial correlations are cut in half (approximately) by using the reduced blocking algorithm which implies that the latter algorithm has higher simulation accuracy (as measured by the numerical standard error) for

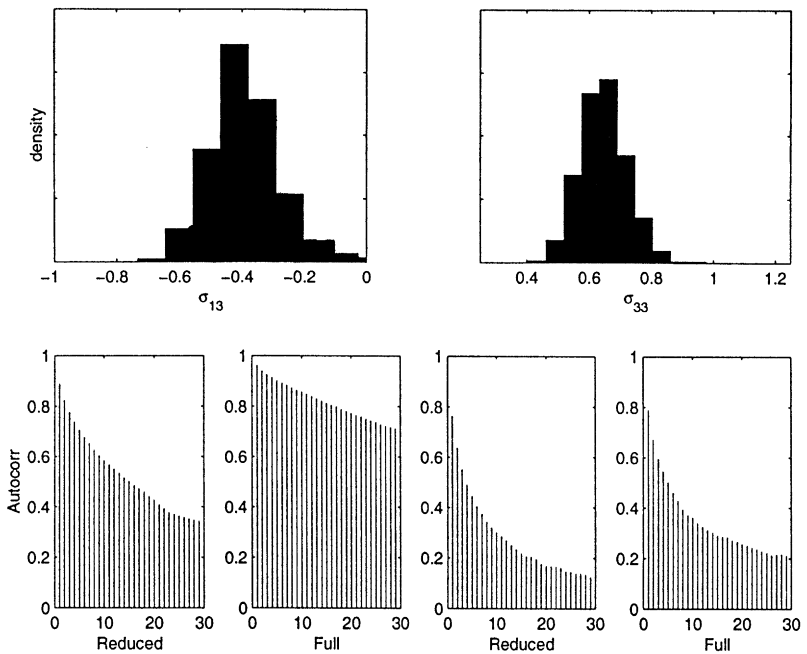

Fig. 1. Results on σ_{13} and σ_{33} in the $t(8)$ model: Posterior distributions (top panel) and autocorrelations in the MCMC output against lag from full and reduced blocking algorithms (bottom panel).

Table 1
Marginal likelihoods of best-fitting models with the Hospice data

	Model		
	Gaussian	$t(8)$	2-Component mixture
Likelihood	− 1027.21	− 991.02	− 981.61
Marginal likelihood	− 1088.53	− 1063.32	− 1060.23

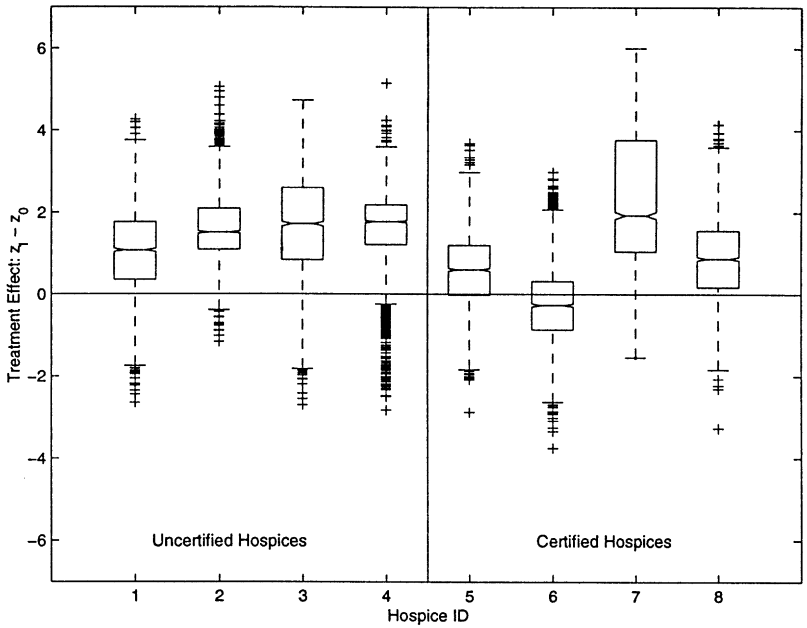

Fig. 2. Posterior distributions of potential outcomes by treatment status for eight randomly selected hospices.

a given Monte Carlo sample size. Note that the posterior distribution of σ_{13} is concentrated on negative values indicating that the unobservables that influence the certification decision are also correlated with the outcome.

Table 1 presents the marginal likelihoods for some of the best-fitting models. The marginal likelihoods do not support the Gaussian model. More support is seen for the Student-t model with eight degrees of freedom, and the two component mixture model. The latter model suggests that the data can be clustered into two groups. In one group, the MHB reimbursement covariate has

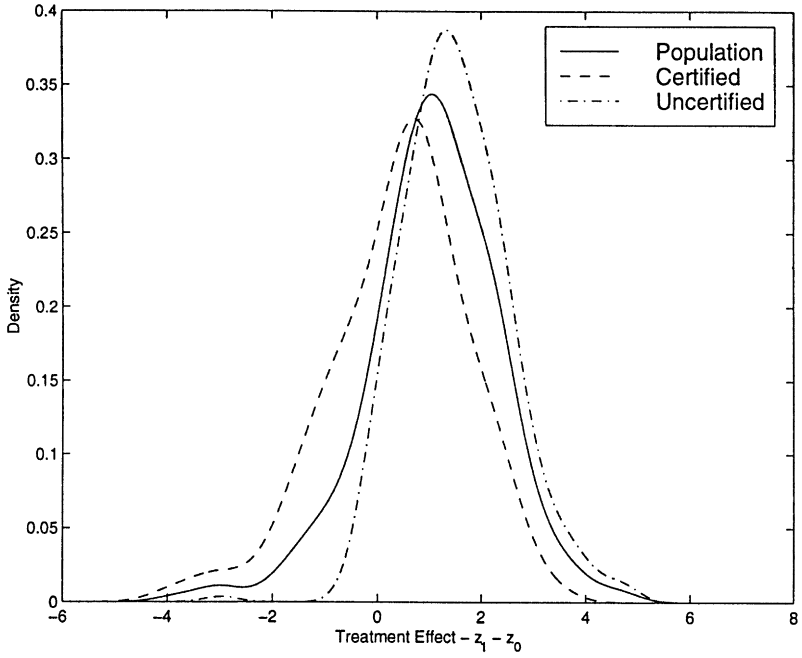

Fig. 3. Posterior densities of treatment effects for entire hospice population and certified and uncertified sub-populations.

a strong positive impact on $\Pr(s_i = 1)$, but in the smaller second group, hospices are insensitive to Medicare reimbursement incentives. Note that there is some preference for the mixture model (the log base ten Bayes factor for the mixture model over the $t(8)$ model is 1.34).

To evaluate the subject level impact of Medicare certification on the access to hospice services, we recorded the sampled values of z_{i0} and z_{i1} from each of the 10,000 iterations of the MCMC sampler for each observation in the sample. Based on these draws, we calculated the hospice-specific distribution of the treatment effect $z_{i1} - z_{i0}$ (which is measured on the log scale). Because it would be cumbersome to present this distribution for all 568 hospices, we randomly selected four uncertified hospices (labeled with ID numbers 1–4) and four certified hospices (labeled with ID numbers 5–8) and plotted summary measures of these distributions in Fig. 2. The boxplots for the four uncertified hospices show that the mean treatment effect is positive and that the bulk of the distribution is supported on positive values. Hospice 4 appears to have the largest mean treatment effect and seems to be the most likely to substantially increase patient capacity if it were to become certified. The treatment effect for

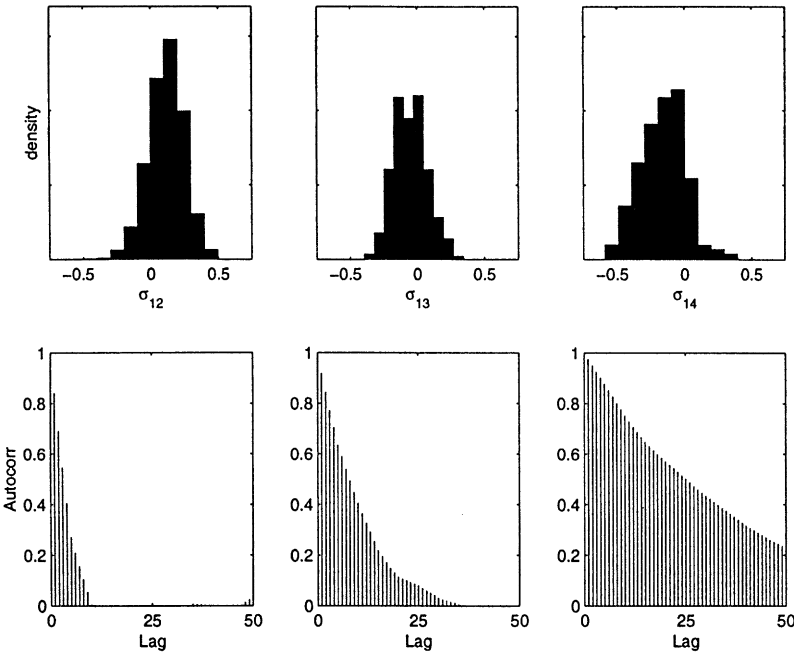

Fig. 4. Results on σ in clustered data model with random effects: posterior distributions (top panel) and autocorrelations in the MCMC output against lag (bottom panel).

hospice 3 appears to be the most uncertain. Among those currently treated (ID 5–8), the treatment effects are generally positive but smaller and more varied than those for hospices 1–4. In the case of hospice 6, the mean treatment effect is negative, while in the case of hospice 7 the mean certification effect is large and the distribution is positively skewed.

As discussed above, the individual hospice treatment effects may be aggregated in a variety of ways or grouped by values of the observed covariates and instruments. The treatment evaluation literature has focussed on the population average treatment effect and, to a lesser extent, on the average treatment effects for the sub-populations observed to receive and not receive the treatment, respectively. Fig. 3 plots the distributions of the mean values of $z_1 - z_0$ for all 568 hospices in the sample (the population average certification effect), as well as for the sub-populations of certified and non-certified hospices. Fig. 4 shows that the population average treatment effect is centered on 1.04 log points, and that the effect of Medicare certification on number of patients served is smaller (though still positive) for the hospices that have chosen certification than for those remaining uncertified.

6. Example 2: Hip fracture data

This example is concerned with the fitting of the ordinal treatment, binary outcome model with clustering that is discussed in Section 3. The variable s_{li} is pre-surgical delay which is categorized into three levels, corresponding to delays of 1–2 days, 3–4 days, and 5 + days, respectively, for a random sample of 2561 female patients who underwent hip fracture surgery in Quebec in 1991 in the lth hospital. Cluster l in this problem is the lth hospital. There are 68 hospitals in the sample. The outcome variable y_i takes the value one if the patient is discharged to home and the value 0 otherwise.

In this problem, the binary potential outcomes are $\{d_{li0}, d_{li1}, d_{li2}\}$, corresponding to $s_{li} = \{0, 1, 2\}$. The associated latent Gaussian potential outcomes are z_{lik}. For the covariates, we let $x_{li0} = x_{li1} = x_{li2}$, where x_{li0} includes age and the number of comorbidities (health conditions such as diabetes or heart trouble) at the time the patient was admitted to the hospital. The covariate vector w_{li} is all of x_{li0} along with an indicator variable representing the day of the week the patient was admitted to the hospital (see Hamilton et al., 1996).

6.1. Model fitting

The model we fit to these data is

$$\begin{pmatrix} s_{li}^{*} \\ z_{li0} \\ z_{li1} \\ z_{li2} \end{pmatrix} \sim \mathcal{N}\left(\begin{pmatrix} w_{li}'\gamma + a_l \\ x_{li0}'\beta_0 + b_{l0} \\ x_{li1}'\beta_1 + b_{l1} \\ x_{li2}'\beta_2 + b_{l2} \end{pmatrix}, \begin{pmatrix} 1 & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{12} & 1 & 0 & 0 \\ \sigma_{13} & 0 & \ddots & \vdots \\ \sigma_{14} & 0 & \cdots & 1 \end{pmatrix} \right), \tag{15}$$

where the random effects $b_l = (a_l, b_{l0}, b_{l1}, b_{l2})$ follow a four-dimensional normal distribution with variance D. The parameters of this model are $\theta = (\xi_1, \sigma, \beta, D)$, where ξ_1 is the single cut-point parameter, $\beta' = (\gamma', \beta_0', \beta_1', \beta_2')$, and $\sigma = (\sigma_{12}, \sigma_{13}, \sigma_{14})$.

We first use the 1990 data with default priors. The results from this training sample estimation are reported in Table 2 under the heading 'training sample posterior'. The posterior mean of θ from this estimation, along with the posterior covariance matrix (suitably inflated), are now used as the parameters of the prior distributions for the 1991 data, as described earlier in the paper. The results from this fit are reported in Table 2 under the heading 'Posterior'. The results are similar across the two sample periods, except for the posterior distribution on D.

In Figs. 4 and 5, we present the posterior histograms of σ and posterior box-plots of D, respectively, and the associated autocorrelation functions of the MCMC output. The distributions of σ_{12} and σ_{14} are mostly concentrated on

Table 2
Posterior estimates for hip fracture data using the clustered model with hospital specific random effects. The second column gives the posterior means and standard deviations from the training sample. The corresponding results for the sample of interest are in the last column. Results are based on 10,000 MCMC draws

Variable	Training sample posterior	Posterior
Treatment (delay)		
Intercept	− 1.427 (0.285)	− 0.857 (0.417)
Age/10	0.037 (0.028)	0.024 (0.028)
Comorbidities	0.088 (0.014)	0.077 (0.013)
Monday	0.080 (0.075)	0.127 (0.073)
Thursday	− 0.106 (0.078)	− 0.200 (0.079)
Potential outcomes (z_{i0})		
Intercept	2.376 (0.639)	2.518 (0.555)
Age/10	− 0.284 (0.075)	− 0.305 (0.057)
Comorbidities	− 0.100 (0.042)	− 0.071 (0.030)
Potential outcomes (z_{i1})		
Intercept	4.886 (2.145)	3.055 (2.707)
Age/10	− 0.502 (0.205)	− 0.296 (0.236)
Comorbidities	− 0.109 (0.103)	− 0.137 (0.086)
Potential outcomes (z_{i2})		
Intercept	2.157 (3.675)	4.783 (4.316)
Age/10	− 0.285 (0.342)	− 0.647 (0.395)
Comorbidities	− 0.188 (0.140)	− 0.242 (0.128)
Cutpoint c	1.102 (0.055)	1.165 (0.094)
σ_{12}	0.160 (0.123)	0.111 (0.132)
σ_{13}	0.001 (0.148)	− 0.051 (0.125)
σ_{14}	− 0.171 (0.152)	− 0.180 (0.165)
D_{11}	2.722 (0.585)	2.350 (0.570)
D_{22}	2.473 (0.601)	3.533 (1.532)
D_{33}	16.570 (9.504)	11.397 (6.601)
D_{44}	15.237 (6.516)	33.616 (27.600)

positive and negative values, respectively, thus indicating that the treatment variable is correlated with the potential outcomes. The posterior distribution of D shows that there is considerable heterogeneity across the clusters. Finally, the mixing of the MCMC output (as measured by the autocorrelations) seems adequate as may be observed from Figs. 4 and 5.

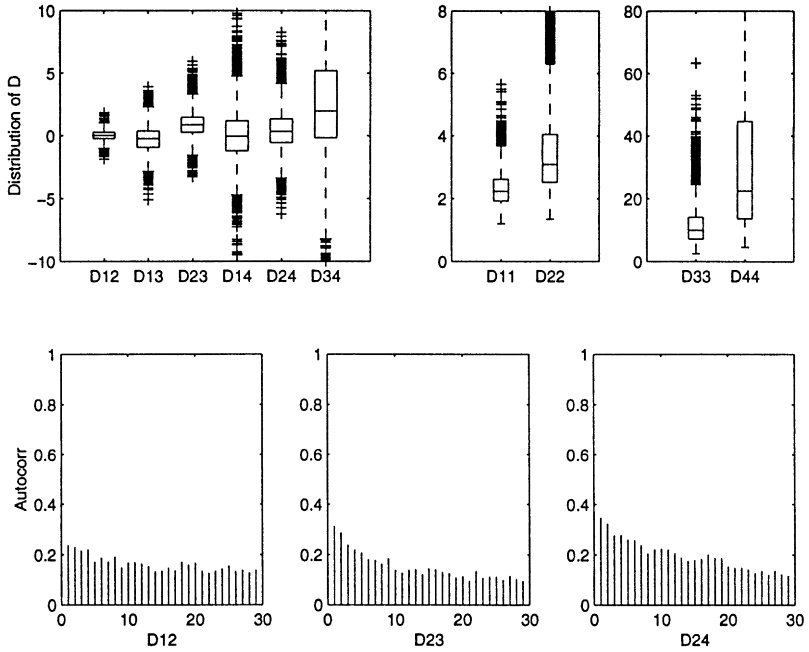

Fig. 5. Results on D in clustered data model with random effects: posterior boxplots (top panel) and selected autocorrelations in the MCMC output against lag (bottom panel).

Table 3
Log likelihood and log marginal likelihood of models without random effects and with random effects

	No random effects	Random effects
$\ln f(y, s \mid \theta^*)$	− 4119.77	− 3969.35
$\ln m(y, s)$	− 4101.03	− 3962.79

Next, based on the prior distributions constructed from our training sample we compare the marginal likelihoods of models with and without random effects. Our results, which are based on 10,000 MCMC draws and a burn-in of 1000 iterations, are presented in Table 3. The table gives the log likelihood and the log marginal likelihood of each model. The marginal likelihoods show that the data strongly support the inclusion of hospital-specific random effects in the model.

6.2. Evaluating the impact of delay on surgical outcomes

To evaluate the impact of surgical delay on the post-surgical probability of a home discharge (from the clustered data model), the first step is to construct the treatment effects for each patient in the sample. As in example 1, we record the sampled values of z_{li0}, z_{li1}, and z_{li2} from each of the 10,000 iterations of the MCMC sampler for each patient in the sample. We then construct the treatment effects associated with a delay of 3–4 days versus 1–2 days, $z_{li1} - z_{li0}$; delay of 5 + days versus 1–2 days, $z_{li2} - z_{li0}$ and delay of 5 + days versus 3–4 days, $z_{li2} - z_{li1}$. Similar treatment effects may be constructed for the binary outcome measures by mapping the z_{lik} into d_{lik} according to (9). Subject-specific treatment effect distributions are plotted in Fig. 6 for four randomly selected patients. The boxplots labeled '10' show that the treatment effect distributions of a delay of 3–4 days versus 1–2 days are centered on zero for three of the four patients. The left most distribution suggests that patient 117, who was delayed for 1–2 days, would have been more likely to be discharged home had she been delayed for 3–4 days. On the other hand, in the case of patient 853 even a delay of 5 + days appears to have no adverse impact on outcomes relative to a shorter delay.

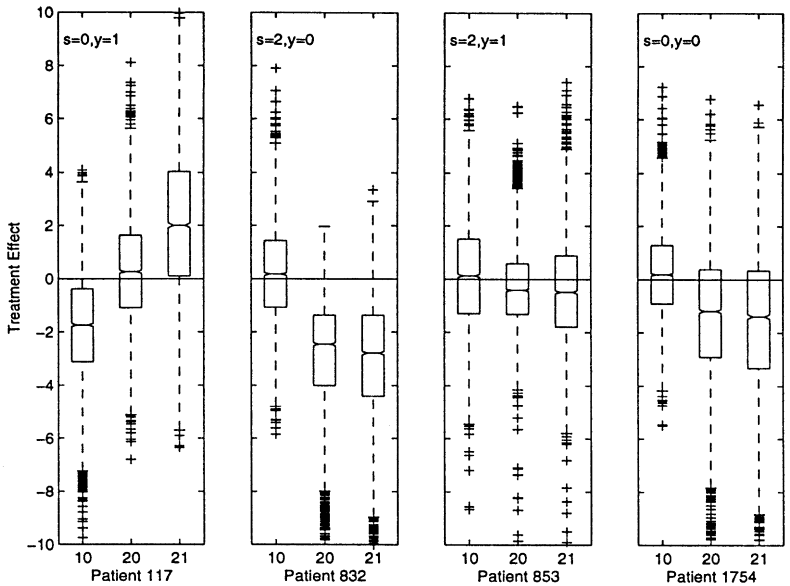

Fig. 6. Posterior distributions of treatment effects for four randomly selected patients. Columns labeled '10' show $z_1 - z_0$; '20' shows $z_2 - z_0$; '21' shows $z_2 - z_1$.

7. Conclusion

This paper has developed a general Bayesian analysis of selection models that incorporate potential outcomes and the non-random assignment of the treatment. We have shown how the Bayesian framework, in conjunction with MCMC methods, can be used to derive the posterior distribution of subject level treatment effects in models that are quite difficult to analyze by any other means. For example, our approach was applied to estimate models in which the potential outcomes follow a mixture of Gaussian distributions. It was also applied to models in which the treatment is ordinal and the responses are binary and clustered. We have discussed the computation of the marginal likelihood with a view to comparing alternative, potentially non-nested selection models and provided evidence in two real data problems for the models that fit well.

We conclude by mentioning that the approach outlined in this paper can be extended in many other useful and important directions, for example in the direction of models with multiple treatments and multivariate (binary, continuous or ordinal) responses, or with more complicated dynamic structures. Work on such models has been initiated and will be reported elsewhere.

Appendix A

Algorithm 1

1. Sample σ from the conditional density $\pi(\sigma|y, s, \beta)$ which is proportional to $g(\sigma) = f_N(\sigma|g_0, G_0)I_S(\sigma)f(y, s|\beta, \Sigma)$, where f_N denotes the kernel of the multivariate normal density. To sample $g(\sigma)$, let $q(\sigma|\mu, V)$ denote a multivariate-t density with parameters μ and V defined as the mode and inverse of the negative Hessian, respectively, of log $g(\sigma)$. Then

(a) Sample a proposal value σ' from the density $q(\sigma|\mu, V)$
(b) Move to σ' given the current point σ with the Metropolis–Hastings probability of move (Chib and Greenberg, 1995)

$$\min\left\{\frac{f_N(\sigma'|g_0, G_0)I_S(\sigma')f(y, s|\beta, \Sigma)}{f_N(\sigma|g_0, G_0)I_S(\sigma)f(y, s|\beta, \Sigma)}\frac{q(\sigma|\mu, V)}{q(\sigma'|\mu, V)}, 1\right\},$$

otherwise stay at σ.

2. Sample z_i^* from $z_i^*|y_i, s_i, \beta, \Sigma$, independently for $i = 1, \ldots, n$. If $s_i = 1$, in which case $z_i^* = (s_i^*, z_{i0})$, first sample s_i^* from the distribution $s_i^*|y_i, s_i, z_{i0}, \beta, \Sigma$, a normal distribution truncated to the interval $(0, \infty)$ and then given this draw of s_i^*, sample z_{i0} from the distribution $z_{i0}|y_i, s_i, s_i^*, \beta, \Sigma$, a normal distribution without any restriction on its support. If $s_i = 0$, then sample s_i^* from the

distribution $s_i^*|y_i, s_i, z_{i1}, \beta, \Sigma$, but now truncated to the interval $(-\infty, 0)$, and then sample z_{i1} from the untruncated normal distribution $z_{i1}|y_i, s_i, s_i^*, \beta, \Sigma$.

3. Sample β from the distribution $\mathcal{N}(\hat{\beta}, B)$ where $\hat{\beta} = B(\beta_0 B_0^{-1} + \sum_{i=1}^n X_i \Sigma^{-1} z_i)$ and $B = (B_0^{-1} + \sum_{i=1}^n X_i \Sigma^{-1} X_i)^{-1}$.

Algorithm 2

1. Sample α and σ from $\alpha, \sigma|y, s, \{b_l\}, \beta$ which is proportional to

$$g(\alpha, \sigma) = \pi(\alpha) f_N(\sigma|g_0, G_0) I_S(\sigma) \prod_{l=1}^{m} f(s_l, y_l|\alpha, \beta, \{b_l\}, \Sigma).$$

To sample $g(\alpha, \sigma)$, let $q(\alpha, \sigma|\mu, V)$ denote a multivariate-t density with parameters μ and V defined as the mode and inverse of the negative Hessian, respectively, of $\log g(\alpha, \sigma)$. Then

(a) Sample a proposal value (α', σ') from the density $q(\alpha, \sigma|\mu, V)$
(b) Move to (α', σ') given the current point (α, σ) with the Metropolis–Hastings probability of move

$$\min\left\{\frac{\pi(\alpha') f_N(\sigma'|g_0, G_0) I_S(\sigma') \prod_{l=1}^m f(s_l, y_l|\alpha', \beta, \{b_l\}, \Sigma')}{\pi(\alpha) f_N(\sigma|g_0, G_0) I_S(\sigma) \prod_{l=1}^m f(s_l, y_l|\alpha, \beta, \{b_l\}, \Sigma)} \frac{q(\alpha, \sigma|\mu, V)}{q(\alpha', \sigma'|\mu, V)}, 1\right\},$$

otherwise stay at (α, σ).

2. Sample $z_{li} = (s_{li}^*, z_{li0}, \ldots, z_{liJ})$ from $z_{li}|s_l, \alpha, \beta, \{b_l\}, \Sigma$, drawing each component of z_{li}. Suppose that the data on the (l,i)th subject is $(s_{li}, y_{li}) = (k, 1)$. Then, sample s_{li}^* from the normal distribution $s_{li}^*|z_{li0}, z_{li1}, \ldots, z_{liJ}, \alpha, \beta, \Sigma$ truncated to the interval (ξ_{k-1}, ξ_k) and then sample z_{lik} (the appropriate potential outcome) from the normal distribution $z_{li1}|s_{li}^*, z_{li0}, z_{li2}, \ldots, z_{liJ}, \alpha, \beta, \Sigma$ truncated to the interval $(0, \infty)$; if y_i is zero, then change the latter interval to $(-\infty, 0)$. Sample the remaining potential outcomes from the appropriate conditional normal distributions but without any restrictions on the support.

3. Sample $\beta = (\gamma, \beta_0, \beta_1, \ldots, \beta_J)$ from $\beta|\{z_{li}\}, \{b_l\}, \Sigma$, where the latter distribution is $\mathcal{N}(\hat{\beta}, B)$, with $\hat{\beta} = B(\beta_0 B_0^{-1} + \sum_{l=1}^m \sum_{i=1}^{n_l} X_{li} \Sigma^{-1}(z_{li} - b_l))$ and

$$B = \left(B_0^{-1} + \sum_{l=1}^{m} \sum_{i=1}^{n} X_{li} \Sigma^{-1} X_{li}\right)^{-1}.$$

4. Sample $b_l = (a_l, b_{l0}, \ldots, b_{lJ})$ from $b_l|\{z_{li}\}, \beta, \Sigma$, where the latter distribution is $\mathcal{N}(\hat{b}_l, C_l)$, with $\hat{b}_l = C_l \Sigma^{-1} \sum_{i=1}^{n_l} (z_{li} - X_{li}\beta)$ and $C_l = (D^{-1} + n_l \Sigma^{-1})$.

5. Sample D^{-1} from $D^{-1}|\{b_l\}$ where

$$D^{-1}|\{b_l\} \sim \text{Wishart}\left(v_0 + n, \left(R_0^{-1} + \sum_{l=1}^{n} n_l (b_l b_l')\right)^{-1}\right)$$

under the assumption that the prior on D^{-1} is Wishart (v_0, R_0).

References

Ahn, H., Powell, J., 1993. Semi-parametric estimation of censored selection models with a non-parametric selection mechanism. Journal of Econometrics 58, 3–29.

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., 1997. Sequential ordinal modeling with applications to survival data. Manuscript.

Angrist, J., Imbens, G., Rubin, D., 1996. Identification of causal effects using instrumental variables (with discussion). Journal of the American Statistical Association 91, 444–455.

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

Chib, S., 1995. Marginal likelihood from the Gibbs output. Journal of the American Statistical Association 90, 1313–1321.

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., 1998. Analysis of multivariate probit models. Biometrika 85, 347–361.

Copas, J., Li, X., 1997. Inference for non-random samples (with discussion). Journal of the Royal Statistical Society B 59, 55–95.

Diebolt, J., Robert, C., 1994. Estimation of finite mixture distributions through Bayesian sampling. Journal of the Royal Statistical Society B 56, 363–375.

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

Hamilton, V., 1993. The Medicare hospice benefit: the effectiveness of price incentives in health care policy. The Rand Journal of Economics 24, 605–624.

Hamilton, B., Hamilton, V., Mayo, N., 1996. What are the costs of queuing for hip fracture surgery in Canada?. Journal of Health Economics 15, 161–185.

Heckman, J., 1978. Dummy endogenous variables in a simultaneous equations system. Econometrica 46, 695–712.

Heckman, J., 1997. Instrumental variables. Journal of Human Resources XXXII, 441–462.

Heckman, J., Robb, R., 1985. Alternative models for evaluating the impact of interventions. In: Heckman, J., Singer, B. (Eds.), Longitudinal Analysis of Labor Market Data. Cambridge University Press, New York, pp. 156–245.

Heckman, J., Smith, J., Taber, C., 1998. Accounting for dropouts in evaluations of social programs. Review of Economics and Statistics LXXX, 1–14.

Holland, P., 1986. Statistics and causal inference. Journal of the American Statistical Association 81, 945–970.

Imbens, G., Rubin, D., 1997. Bayesian inference for causal effects in randomized experiments with noncompliance. Annals of Statistics 25, 305–327.

Koop, G., Poirier, D., 1997. Learning about the across-regime correlation in switching regression models. Journal of Econometrics 78, 217–227.

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

Lee, L.F., 1979. Identification and estimation in binary choice models with limited (censored) dependent variables. Econometrica 47, 977–996.

Maddala, G., 1983. Limited-Dependent and Qualitative Variables in Econometrics. Cambridge University Press, Cambridge, UK.

Newey, W., Powell, J., Walker, J., 1990. Semiparametric estimation of selection models: some empirical results. American Economic Review — Papers and Proceedings 80, 324–328.

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., 1987. The calculation of posterior distributions by data augmentation (with discussion). Journal of the American Statistical Association 82, 528–550.

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

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