Available online at www.sciencedirect.com
ScienceDirect
Journal of Econometrics 140 (2007) 401–412
JOURNAL OF Econometrics
www.elsevier.com/locate/jeconom

# Analysis of treatment response data without the joint distribution of potential outcomes

Siddhartha Chib*

*John M. Olin School of Business, Washington University in St. Louis, CB 1133, 1 Brookings Dr., St. Louis, MO 63130, USA*

Available online 22 August 2006

**Abstract**

In this paper we show how it is possible to develop a Bayesian framework for analyzing structural models for treatment response data without the joint distribution of the potential outcomes. That this is possible has not been noticed in the literature. We also discuss the computation of the model marginal likelihood and present recipes for finding relevant treatment effects, averaged over both parameters and covariates. As compared to an approach in which the counterfactuals are part of the prior-posterior analysis (as in the work to date), the approach we suggest is simpler in terms of the required prior inputs, computational burden and extensibility to more complex settings.


*Keywords:* Confounding; Instrumental variable; Marginal likelihood; Markov chain Monte Carlo; Structural model; Predictive treatment effect

## 1. Introduction

The structural potential outcomes regression model is one of the most useful models for the analysis of treatment–response data. The simplest model of this type, which was introduced by Lee (1978) for the case of a continuous response and a binary treatment, postulates a marginal model for the outcomes under each treatment state depending on covariates and treatment specific parameters, a marginal model for the treatment, which may or not depend on an additional set of covariates (namely instruments), and unobserved confounders that affect both the response and the treatment. The basic model

*Tel.: +1 314 935 6359.
*E-mail address:* chib@wustl.edu.

0304-4076/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.jeconom.2006.07.009

has been subjected to considerable scrutiny from both the frequentist and Bayesian perspectives. Analysis of the model in the former context is generally based on the likelihood function, and provided one takes the usual care in formulating the distributional assumptions, the likelihood function is easily obtained, at least for the basic model (for example, Amemiya, 1985, p. 400), and analysis is straightforward.

To understand the existing Bayesian approaches for the basic structural potential outcomes model it is helpful to recall that in the context of treatment–response problems with a binary treatment variable, say denoted by $x \in \{0, 1\}$, there are two potential outcomes, $y_x$. For each observation $i$, only one of these potential outcomes is observed, the other potential outcome is the counterfactual. This feature implies that the joint distribution of the potential outcomes is not identified. Nonetheless, motivated by the missingness of the counterfactual, Vijverberg (1993) and Poirier and Koop (1997) formulated a Bayesian analysis for this model with the joint distribution of the potential outcomes. Putting aside the details of the distributional assumptions, estimation under their framework requires a prior on the non-identified covariance parameter of the joint distribution. The model is then estimated by MCMC methods by simulating the posterior distribution of the parameters *and* the counterfactuals. Subsequently, within the context of the joint modeling of the potential outcomes, Chib and Hamilton (2000, 2002) provided a Bayesian analysis of generalized versions of the basic model relevant for panel data, binary outcomes and ordinal treatments, under weak distributional assumptions, while Poirier and Tobias (2003) revisited the basic model with a prior on the non-identified covariance parameter that was different than that in Chib and Hamilton (2000, 2002).

The goal of the current paper is to provide a Bayesian analysis of the basic structural potential outcomes model without the involvement of a joint model of the potential outcomes. That this is possible has not been noticed in the literature. Because direct use of the likelihood function does not lead to a tractable posterior distribution, and because the likelihood function is not easily available in generalized versions of the basic model, our approach which we detail below, does not involve the likelihood function directly, or the missing counterfactuals. Still, by taking advantage of the framework of Albert and Chib (1993), we obtain a target posterior distribution that can be processed readily by MCMC methods. Additionally, because the analysis is free of unnecessary counterfactuals and unidentified parameters, the new approach is actually simpler in terms of the required prior inputs, computational burden and extensibility to more complex settings. It may be mentioned that the issues discussed in this paper are distinct from those in Dawid (2000, 2003), who has made a case against involving counterfactuals in causal problems but has not discussed how his approach would be operationalized in the sort of model we consider in this paper.

To complete our inferential approach, we also discuss the questions of model comparisons across competing models and the computation of various treatment effects from a predictive framework. One of the treatment effects we consider is the treatment effect for compliers, similar to Imbens and Angrist (1994) in a different context. This treatment effect depends on the notion of potential treatments (the treatment under each level of the instrument). An interesting point is that this effect can be calculated from the marginal distribution of the treatment without recourse to a joint distribution of the potential treatments. Thus, in the framework we develop, estimation proceeds without the joint distribution of the potential outcomes, and the computation of the treatment effect without the joint distribution of potential treatments. It also worth pointing out that from

our framework it is not possible to calculate the variability of the treatment effect because this computation requires the joint distribution of the potential outcomes. This is not a drawback, however, because one can reasonably argue that unidentified quantities (such as the variability in the treatment effect) should not be objects of interest in the first place.

The remainder of the paper is organized as follows. In Section 2 we present the model and the prior distribution for the parameters. In Section 3 we describe the fitting of the model by a tuned MCMC method and discuss the computation of the marginal likelihood from the MCMC output. Section 4 deals with the computation of the treatment effect, Section 5 has results from simulation studies and Section 6 concludes.

## 2. Model and prior distribution

For each subject $i$ in the sample, let $x_i \in \{0, 1\}$ denote the binary treatment indicator, where $x_i$ is not randomly assigned, and let $y_{0i}$ and $y_{1i}$ denote the corresponding potential outcomes. The observed response is

$$y_i = y_{0i} + (y_{1i} - y_{0i})x_i \tag{2.1}$$

which is $y_{0i}$ if the treatment is not received or $y_{1i}$, otherwise. Let $\mathbf{w}_i \in R^k$ denote a vector of covariates and let $u_i$ denote an unobserved random variable. Assume that the effect of $x_i$ on the response is confounded with that of $u_i$, conditional on the covariates. Furthermore, suppose that there is a covariate $z_i$ (an instrument) that is correlated with the treatment but uncorrelated with $(y_{0i}, y_{1i}, u_i)$ given the covariates.

To model the potential outcomes and the treatment assume that

$$\begin{aligned} y_{ji} &= \mathbf{w}_i'\boldsymbol{\beta}_j + \varepsilon_{ji}, \quad j = 0, 1, \\ x_i^* &= \mathbf{w}_i'\boldsymbol{\gamma} + z_i\delta + u_i, \\ x_i &= I\{x_i^* > 0\}, \end{aligned} \tag{2.2}$$

where $\boldsymbol{\beta}_j \in R^k$, $\boldsymbol{\gamma} \in R^k$ and $\delta$ are unknown parameters, and $I(\cdot)$ is the indicator function. In this specification, the first equation (as $j$ takes the values 0 and 1) generates the marginal distribution of the potential outcomes and the second and third generate the marginal distribution of the treatment.

The next vital step in the modeling is the specification of the needed joint distributions of the outcomes and the intake. In formulating these distributions we have to contend with the fact that the joint distribution of the potential outcomes is unidentified because the outcomes $y_{0,i}$ and $y_{1,i}$ cannot be observed simultaneously. The practice to date has been to assume some specific but unverifiable form for this joint distribution. This practice is rather unsatisfactory since the analysis then involves the unidentified parameters of that joint distribution and the missing counterfactuals (one for each subject). We resolve this difficulty by noticing that because the responses for the $i$th subject are either $(y_{0i}, x_i = 0)$ or $(y_{1i}, x_i = 1)$, the modeling can be completed by simply specifying the bivariate joint distributions $(\varepsilon_{0i}, u_i)$ and $(\varepsilon_{1i}, u_i)$ or, equivalently, the joint distributions $p_0(y_{0i}, x_i = 0|\mathbf{w}_i, z_i)$ and $p_1(y_{1i}, x_i = 1|\mathbf{w}_i, z_i)$, such that the marginal distribution of $u_i$ is the same in each case.

To illustrate the main themes, assume for specificity that $(\varepsilon_{ji}, u_i)$ given $\lambda_i$ is

$$(\varepsilon_{ji}, u_i)|\lambda_i \sim \mathrm{N}_2(\mathbf{0}, \lambda_i^{-1}\boldsymbol{\Omega}_j),$$

where $\lambda_i$ is a positive random-variable that is iid gamma $(v/2, v/2)$ for some known value $v>0$, and

$$\mathbf{\Omega}_j = \begin{pmatrix} \eta_j^2 & \omega_j \\ \omega_j & 1 \end{pmatrix}.$$

Thus, $\lambda_i^{-1}\mathbf{\Omega}_j$ is the conditional covariance matrix between the $j$th potential outcome and the treatment (on the latent scale $x_i^*$) and, marginally of $\lambda_i$, the joint distribution of the treatment and the outcome is student-$t$.

In anticipation of the estimation procedure we develop in the sequel, we note that the parameterization of $\mathbf{\Omega}_j$ by means of $\eta_j$ and $\omega_j$ is not convenient because these parameters must satisfy the positive-definiteness constraint. It is helpful instead to work with the parameters $\sigma_j^2 = \eta_j^2 - \omega_j^2$, which is the determinant of $\mathbf{\Omega}_j$, and $\omega_j$.

Let $\boldsymbol{\psi}_j = (\sigma_j^2, \omega_j)$ and $\boldsymbol{\beta} = (\boldsymbol{\beta}_0, \boldsymbol{\beta}_1, \boldsymbol{\gamma}, \delta)$. Then the assumptions stated above imply that the joint density of $(y_{ji}, x_i^*)$ conditioned on the parameters is

$$p_j^*(y_{ji}, x_i^* | \mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_j) = \mathrm{N}_2(y_{ji}, x_i^* | \mathbf{X}_{ji}\boldsymbol{\beta}, \lambda_i^{-1}\mathbf{\Omega}_j), \tag{2.3}$$

where

$$\mathbf{X}_{ji} = \begin{pmatrix} \mathbf{w}_i' \times (1-j) & \mathbf{w}_i' \times j & \mathbf{0}' & 0 \\ \mathbf{0}' & \mathbf{0}' & \mathbf{w}_i' & z_i \end{pmatrix}.$$

From here the contribution $p_j(y_{ji}, x_i = j | \mathbf{w}_i, z_i, \boldsymbol{\beta}, \boldsymbol{\psi}_j)$ of the $i$th observation to the likelihood can be derived by integrating out $x_i^*$. Specifically, by utilizing the properties of the multivariate-$t$ distribution (see for example, Bilodeau and Brenner, 1999, p. 239) it follows that

$$\begin{aligned} p_j(y_{ji}, x_i = j | \mathbf{w}_i, z_i, \boldsymbol{\beta}, \boldsymbol{\psi}_j) &= p_j(y_{ji} | \mathbf{w}_i, \boldsymbol{\beta}, \boldsymbol{\psi}_j) \int_{A_j} p_j^*(x_i^* | \mathbf{w}_i, z_i, y_{ji}, \boldsymbol{\beta}, \boldsymbol{\psi}_j)\, \mathrm{d}x_i^* \\ &= t_v(y_{ji} | \mathbf{w}_i'\boldsymbol{\beta}_j, \eta_j^2) T_{v+1}\left((2j-1)\frac{\mu_{ji}}{h_{ji}\phi_j}\right), \end{aligned} \tag{2.4}$$

where $A_j$ is the set $(-\infty, 0)$ if $j=0$ or $(0, \infty)$ if $j=1$, $t_v(.|\mu, \sigma^2)$ is the density of the student-$t$ density with $v$ degrees of freedom, location parameter $\mu$ and dispersion parameter $\sigma^2$, $T_{v+1}$ is the cdf of the $t_v(.|0,1)$ density, $\mu_{ji} = \mathbf{w}_i'\boldsymbol{\gamma} + z_i\delta + \omega_j\eta_j^{-2}(y_{ji} - \mathbf{w}_i'\boldsymbol{\beta}_j)$, $h_{ji}^2 = [\upsilon(\upsilon+1)][1 + (y_{ji} - \mathbf{w}_i'\boldsymbol{\beta}_j)^2\eta_j^{-2}/v]$ and $\phi_j^2 = 1 - \omega_j^2/\eta_j^2$.

### 2.1. Prior distribution

Our approach to inference is Bayesian so we complete the model specification by defining the prior distribution of the model parameters $\boldsymbol{\beta}$ and $\boldsymbol{\psi}_j$ $(j=0,1)$. Following conventional practice for dealing with regression parameters, we assume that $\boldsymbol{\beta}$ is a priori $\mathrm{N}_p(\boldsymbol{\beta}|\mathbf{b}_0, \mathbf{B}_0)$, where $p$ is the dimension of $\boldsymbol{\beta}$, and the quantities indexed by zero are known hyperparameters. Next, we model $\boldsymbol{\psi}_j$ jointly by assuming that $\sigma_j^2$ (which must be positive) is inverse-gamma, and that $\omega_j$ conditioned on $\sigma_j^2$ is Gaussian:

$$\pi(\boldsymbol{\psi}_j) = \text{inverse gamma}\left(\sigma_j^2 \middle| \frac{n_{j0}}{2}, \frac{d_{j0}}{2}\right) \mathrm{N}(\omega_j | m_{j,0}, \sigma_j^2 M_{j,0}).$$

With the further assumption that the different blocks of parameters are a priori independent our prior density is of the form

$$\pi(\boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1) = \mathrm{N}_p(\boldsymbol{\beta}|\mathbf{b}_0, \mathbf{B}_0) \prod_{j=0}^{1} \text{inverse gamma}\left(\sigma_j^2 \Big| \frac{n_{j0}}{2}, \frac{d_{j0}}{2}\right) \mathrm{N}(\omega_j | m_{j,0}, \sigma_j^2 M_{j,0}). \quad (2.5)$$

Finally, as stated above, the remaining unknowns in the model, the $\lambda_i$'s, are modeled as independent gamma $(\nu/2, \nu/2)$.

## 3. Estimation and model comparison

### 3.1. Estimation

Suppose now that we have a random sample of responses $(y_1, x_1), \ldots, (y_n, x_n)$ on $n$ subjects. The goal is to learn about the parameters $(\boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ given the data, the model and our prior inputs. Clearly, prior-posterior analysis with the likelihood function of the data is not convenient. It is possible, however, to develop a tractable approach that does not involve the likelihood function directly or the missing counterfactuals. To do this, we take advantage of the approach of Albert and Chib (1993) and operate with the conditional densities

$$p(y_i, x_i^*, x_i = 0|\mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_0) = p_0^*(y_i, x_i^*|\mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_0) I\{x_i^* < 0\},$$
$$p(y_i, x_i^*, x_i = 1|\mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_1) = p_1^*(y_i, x_i^*|\mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_1) I\{x_i^* > 0\},$$

where $p_j^*(y_i, x_i^*|\mathbf{w}_i, z_i, \lambda_i, \boldsymbol{\beta}, \boldsymbol{\psi}_j) = \mathrm{N}_2(y_i, x_i^*|\mathbf{X}_{ji}\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Omega}_j)$ is the bivariate normal density from (2.3). The posterior distribution of interest is then

$$\pi(\boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\lambda}, \mathbf{x}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) \propto \pi(\boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)\pi(\boldsymbol{\lambda}) \prod_{i \in N_0} \mathrm{N}_2(y_i, x_i^*|\mathbf{X}_{0i}\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Omega}_0) I\{x_i^* < 0\}$$
$$\times \prod_{i \in N_1} \mathrm{N}_2(y_i, x_i^*|\mathbf{X}_{1i}\boldsymbol{\beta}, \lambda_i^{-1}\boldsymbol{\Omega}_1) I\{x_i^* > 0\}, \quad (3.1)$$

where $\boldsymbol{\lambda} = \{\lambda_i\}$, $\mathbf{x}^* = (x_1^*, \ldots, x_n^*)$, $(\mathbf{W}, \mathbf{z})$ is the data on covariates and instruments, and $N_j = \{i\colon\ x_i = j\}, j = 0, 1$. This joint distribution, which is free of the missing counterfactuals, is of a type that can be sampled by MCMC methods (Tierney, 1994; Chib and Greenberg, 1995). In particular, we can design a 3-block sampling scheme that is both easy to implement and simulated efficient.

In the first block of the scheme, $\boldsymbol{\psi}_j = (\sigma_j^2, \omega_j)$ are sampled jointly by the method of composition. To see how, note that the distribution of $y_{ji}$ conditional on $x_i^*$ and everything else is

$$y_{ji}|x_i^*, \boldsymbol{\beta}_j, \lambda_i, \boldsymbol{\psi}_j \sim \mathrm{N}(\mathbf{w}_i'\boldsymbol{\beta}_j + \omega_j u_i, \lambda_i^{-1}\sigma_j^2),$$

where $u_i = x_i^* - \mathbf{w}_i'\boldsymbol{\gamma} - z_i\delta$. Now let

$$\mathbf{y}_j = \{y_{ji}\}, \mathbf{x}_j^* = \{x_i^*\}, \mathbf{u}_j = \{x_i^* - \mathbf{w}_i'\boldsymbol{\gamma} - z_i\delta\},\ i \in N_j$$

denote vectors of dimension $n_j \times 1$ obtained by assembling observations that are in $N_j$. From those same observations, let $\mathbf{W}_j = \{\mathbf{w}_i'\}$ denote a stacked $n_j \times k$ matrix, and $\boldsymbol{\Lambda}_j = \mathrm{diag}\{\lambda_i\}$ denote the matrix with $\lambda_i$'s on the diagonal. We then have that

$$\mathbf{y}_j|\mathbf{x}_j^*, \boldsymbol{\beta}_j, \boldsymbol{\Lambda}_j, \boldsymbol{\psi}_j \sim \mathrm{N}_{n_j}(\mathbf{W}_j\boldsymbol{\beta}_j + \omega_j\mathbf{u}_j, \sigma_j^2\boldsymbol{\Lambda}_j^{-1}).$$

This distribution leads easily to the distribution of $\omega_j$ conditioned on $\sigma_j^2$. Furthermore, marginalized over $\omega_j$ under the prior $\omega_j \sim \mathrm{N}(m_{j,0}, \sigma_j^2 M_{j,0})$ we also get that

$$\mathbf{y}_j | \mathbf{x}_j^*, \boldsymbol{\beta}_j, \boldsymbol{\Lambda}_j, \sigma_j^2 \sim \mathrm{N}_{n_j}(\mathbf{W}_j \boldsymbol{\beta}_j + m_{j,0}\mathbf{u}_j, \sigma_j^2(\boldsymbol{\Lambda}_j^{-1} + \mathbf{u}_j M_{j,0} \mathbf{u}_j')).$$

This can now be combined with the prior on $\sigma_j^2$ to produce the distribution of $\sigma_j^2$ marginalized over $\omega_j$. In the next step, $\boldsymbol{\beta}$ is sampled conditioned on everything else. Finally in the third block, $(\mathbf{x}^*, \boldsymbol{\lambda})$ are sampled jointly, again by the method of composition. In detail, we have the following MCMC algorithm for sampling the posterior distribution in (3.6):

1. Sample $\boldsymbol{\psi}_j = (\sigma_j^2, \omega_j)$, $j = 0, 1$ conditioned on $(\mathbf{y}, \mathbf{x}^*, \mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\lambda})$ by
   (a) drawing $\sigma_j^2$ marginalized over $\omega_j$ from
   $$\text{inverse gamma}\left(\frac{v_{j,0} + n_j}{2}, \frac{d_{j,0} + d_j}{2}\right),$$
   where $d_j = (\mathbf{e}_j - m_{j,0}\mathbf{u}_j)'(\boldsymbol{\Lambda}_j^{-1} + \mathbf{u}_j M_{j,0} \mathbf{u}_j')^{-1}(\mathbf{e}_j - m_{j,0}\mathbf{u}_j)$, and $\mathbf{e}_j = \mathbf{y}_j - \mathbf{W}_j \boldsymbol{\beta}_j$
   (b) drawing $\omega_j$ conditioned on $\sigma_j^2$ from
   $$\mathrm{N}(b_j, \sigma_j^2 B_j),$$
   where $b_j = B_j(M_{j,0}^{-1} m_{j,0} + \mathbf{u}_j' \boldsymbol{\Lambda}_j \mathbf{e}_j)$ and $B_j = (M_{j,0}^{-1} + \mathbf{u}_j' \boldsymbol{\Lambda}_j \mathbf{u}_j)^{-1}$
2. Sample $\boldsymbol{\beta}$ conditioned on $(\mathbf{y}, \mathbf{x}^*, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1, \boldsymbol{\lambda})$ from
   $$\mathrm{N}_p(\hat{\boldsymbol{\beta}}, \mathbf{B}),$$
   where $\hat{\boldsymbol{\beta}} = \mathbf{B}\left(\mathbf{B}_0^{-1}\mathbf{b}_0 + \mathbf{A}_0 + \mathbf{A}_1\right)$, $\mathbf{A}_j = \sum_{i \in N_j} \lambda_i \mathbf{X}_{ji}' \boldsymbol{\Omega}_j^{-1} \mathbf{y}_i^*$, $\mathbf{y}_i^* = (y_i, x_i^*)'$ and
   $$\mathbf{B} = \left(\mathbf{B}_0^{-1} + \sum_{i \in N_0} \lambda_i \mathbf{X}_{0i}' \boldsymbol{\Omega}_0^{-1} \mathbf{y}_i^* + \sum_{i \in N_1} \lambda_i \mathbf{X}_{1i}' \boldsymbol{\Omega}_1^{-1} \mathbf{y}_i^*\right)^{-1}.$$
3. Sample $(\mathbf{x}^*, \boldsymbol{\lambda})$ conditioned on $(\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}, \boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ by
   (a) drawing $x_i^*$ from $t_{v+1}(\mu_{0i}, \phi_0^2) I_{(-\infty,0)}$ if $x_i = 0$ and from $t_{v+1}(\mu_{1i}, \phi_1^2) I_{(0,\infty)}$ if $x_i = 1$ $(i \leqslant n)$;
   (b) drawing $\lambda_i$ $(i \leqslant n)$ conditioned on $(\mathbf{y}_i^*, x_i, \boldsymbol{\beta}, \boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ from
   $$\text{gamma}\left(\frac{v+2}{2}, \frac{v + (\mathbf{y}_i^* - \mathbf{X}_{ji}\boldsymbol{\beta})' \boldsymbol{\Omega}_j^{-1} (\mathbf{y}_i^* - \mathbf{X}_{ji}\boldsymbol{\beta})}{2}\right).$$

(4) Go to 1.

It is easy to see from the description of the algorithm that the sampling steps are straightforward. Thus, in contrast to what occurs in other problems, the non-inclusion of the missing data (here the unobserved counterfactuals) has no bearing on the complexity of the fitting procedure. Another point to note is that the approach above can be extended relatively easily to more complicated situations, for example, time-varying treatments as in a panel context. Such an extension would be less straightforward if the joint distribution of the potential outcomes was part of the modeling and posterior sampling. To see this,

consider the case of a model with $J+1$ ordinal treatments that is discussed in Chib and Hamilton (2000). Then, there are $J+1$ potential outcomes and the counterfactuals approach requires the joint distribution of $(\varepsilon_{0i}, \ldots, \varepsilon_{Ji}, u_i)$ with dispersion matrix

$$\mathbf{\Omega} = \begin{pmatrix} \eta_0^2 & \xi_{01} & \cdots & \xi_{0J} & \omega_0 \\ \xi_{01} & \eta_1^2 & \cdots & \xi_{1J} & \omega_1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \xi_{0J} & \xi_{1J} & \cdots & \eta_J^2 & \omega_j \\ \omega_0 & \omega_1 & \cdots & \omega_j & 1 \end{pmatrix},$$

where all the $(J+1)J/2\xi_{jk}$'s in the upper $(J+1)\times(J+1)$ sub-block of this matrix are unidentified. The analysis is now more complex because of the need to specify a prior on these unidentified parameters, and the involvement of not only these unidentified parameters in the sampling but also the $J$ counterfactual variables for each subject. In contrast, in the extension of the framework we have proposed, analysis would be based on the $J+1$ joint distributions $(\varepsilon_{ji}, u_i)$, free of the $\xi_{jk}$'s and the unobserved counterfactuals.

### 3.2. Model comparison

In practice one would be interested in gauging the support for a given model of the type we have just fit against one or more competing models (say defined through a different set of covariates or without confounding on unobservables). In accordance with formal Bayesian precepts the relative support for the contending models can be computed in terms of the pairwise Bayes factors, obtained as ratios of marginal likelihoods. The marginal likelihood of the model above is easily computed by the method of Chib (1995). The basic idea is that on the log-scale the marginal likelihood $m(\mathbf{y}, \mathbf{x}|\mathbf{W}, \mathbf{z})$ can be written as

$$\begin{aligned} \ln m(\mathbf{y}, \mathbf{x}|\mathbf{W}, \mathbf{z}) = & \ln f(\mathbf{y}, \mathbf{x}|\mathbf{W}, \mathbf{z}, \boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*) + \ln \pi(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*) \\ & - \ln \pi(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}), \end{aligned}$$

where $(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*)$ is (say) the posterior mean of the parameters from the MCMC run, the first term is the log likelihood, the second is the prior, and the third is the posterior, each evaluated at $(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*)$. The first two terms are clearly available directly. For example, the first term is given by

$$\sum_{i \in N_0} \ln p_0(y_{ji}, x_i = 0|\mathbf{w}_i, z_i, \boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*) + \sum_{i \in N_1} \ln p_1(y_{ji}, x_i = 1|\mathbf{w}_i, z_i, \boldsymbol{\beta}^*, \boldsymbol{\psi}_1^*),$$

where $p_j(y_{ji}, x_i = j|\mathbf{w}_i, z_i, \boldsymbol{\beta}^*, \boldsymbol{\psi}_j^*)$ appears in (2.4). The third can be estimated efficiently by decomposing it as

$$\pi(\boldsymbol{\beta}^*, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) = \pi(\boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) \pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*),$$

where $\pi(\boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$ is obtained by averaging the product of the inverse gamma and normal densities in Step 1 of the MCMC algorithm over the MCMC draws; and $\pi(\boldsymbol{\beta}^*|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}, \boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*)$ is obtained by fixing $(\boldsymbol{\psi}_0, \boldsymbol{\psi}_1)$ at $(\boldsymbol{\psi}_0^*, \boldsymbol{\psi}_1^*)$, running the MCMC algorithm with the remaining unknowns and averaging the normal density in Step 2 of the MCMC algorithm over the resulting draws.

## 4. Inferring treatment effects

We now describe how the output of the MCMC fitting algorithm can be used to infer two useful treatment effects of interest. A particular treatment effect parameter is the so-called average treatment effect (ATE) which is defined as the difference in the marginal means of the potential outcomes:

$$\text{ATE} = \mathbb{E}(y_1) - \mathbb{E}(y_0).$$

This parameter measures the effect of an intervention on $x$ and involves consideration of only the marginal distribution of the potential outcomes because an intervention severs any link to the probability model that determines the treatment assignment (see for example, Pearl, 2000).

In our set-up, instead of directly starting with the ATE, it is possible to consider an alternative ATE type effect that is based on a predictive approach. Consider a new subject $n+1$ drawn randomly from the population. Our aim is to calculate the marginal density of each potential outcome $y_{j,n+1}$ $(j = 0, 1)$ without reference to the treatment assignment model. Each of these marginal densities is obtained as

$$p(y_{j,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) = \int p(y_{j,n+1}|\mathbf{w}_{n+1}, \lambda_{n+1}, \boldsymbol{\beta}_j, \eta_j^2) \times \pi(\mathbf{w}_{n+1}, \lambda_{n+1}, \boldsymbol{\lambda}, \boldsymbol{\beta}_j, \eta_j^2|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})\, \mathrm{d}\lambda_{n+1}\, \mathrm{d}\boldsymbol{\lambda}\, \mathrm{d}\boldsymbol{\beta}_j\, \mathrm{d}\eta_j^2\, \mathrm{d}\mathbf{w}_{n+1}, \tag{4.1}$$

where $p(y_{j,n+1}|\mathbf{w}_{n+1}, \lambda_{n+1}, \boldsymbol{\beta}_j, \eta_j^2)$ is $\mathrm{N}(y_{j,n+1}|\mathbf{w}'_{n+1}\boldsymbol{\beta}_j, \eta_j^2)$, the marginal density of the potential outcome of subject $n+1$ conditioned on the parameters; this density does not depend on the sample $\mathbf{y}$ or $\mathbf{x}$ because subject $n+1$ is randomly drawn from the population. In this calculation, by the usual rules of probability, the unknowns $(\mathbf{w}_{n+1}, \lambda_{n+1}, \boldsymbol{\lambda}, \boldsymbol{\beta}_j, \eta_j^2)$ are marginalized with respect to the posterior distribution (the marginal posterior distribution of $\mathbf{w}_{n+1}$ may be approximated by the empirical distribution of the covariates given the current sample $\mathbf{W}$). Although the integration cannot be performed analytically it is a simple matter to obtain a sample of draws from $p(y_{j,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$ by the method of composition. In particular, at the $g$th $(g \leqslant M)$ iteration of our MCMC algorithm, when the current state of the chain is defined by the quantities $(\boldsymbol{\lambda}^{(g)}, \boldsymbol{\beta}_j^{(g)}, \eta_j^{2(g)})$, we get a draw $y_{j,n+1}^{(g)}$ from the predictive distribution by using the following steps:

- Sample $\mathbf{w}_{n+1}^{(g)}$ by assigning probability $1/n$ to each row of $\mathbf{W}$.
- Sample $\lambda_{n+1}^{(g)}$ from gamma $(v/2, v/2)$.
- Sample $y_{j,n+1}^{(g)}$ from $\mathrm{N}(\mathbf{w}_{n+1}^{(g)'}\boldsymbol{\beta}_j^{(g)}, \eta_j^{2(g)}/\lambda^{(g)})$.

This gives rise to the desired sample $\{y_{j,n+1}^{(1)}, \ldots, y_{j,n+1}^{(M)}\}$ from the predictive distribution. In the usual way, the expected value of $y_{j,n+1}$ under the distribution $p(y_{j,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})$ can be calculated as

$$\mathbb{E}(y_{j,n+1}|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z}) = \int y_{j,n+1} p(y_j|\mathbf{y}, \mathbf{x}, \mathbf{W}, \mathbf{z})\, \mathrm{d}y_j \simeq \frac{1}{M}\sum_{g=1}^{M} y_{j,n+1}^{(g)}, \tag{4.2}$$

with the difference $\mathbb{E}(y_{1,n+1}|\mathbf{y},\mathbf{x},\mathbf{W},\mathbf{z}) - \mathbb{E}(y_{0,n+1}|\mathbf{y},\mathbf{x},\mathbf{W},\mathbf{z})$ being (what we might call) the predictive ATE. More interestingly, given the draws from each predictive distribution, we can compare quantities other than the mean. For example, we can consider the difference in the specified quantiles of each predictive distribution, estimated by the difference in the corresponding sample quantiles of the sampled draws.

Another treatment effect is defined by thinking in terms of the marginal density of the potential outcomes for the sub-group of subjects who are compliers. This question was considered in a different set-up by Imbens and Angrist (1994). For simplicity, suppose that the instrument is a binary $\{0,1\}$ variable, which is the typical case in practice. Now let $x_0$ denote the treatment when $z=0$ and $x_1$ denote the treatment when $z=1$; these are the potential treatments, only one of which is observed depending on the value of $z$. When $\delta>0$ we say an individual is a complier if $x_{0,n+1}=0$ *and* $x_{1,n+1}=1$. Likewise if $\delta<0$, a complier is an individual for whom $x_{0,n+1}=1$ *and* $x_{1,n+1}=0$. An important point is that one can make predictive inferences about compliance from the marginal distribution of the treatment given in (2.2) without involvement of the unidentified joint distribution of the potential treatments. The basic idea is to set $z_{n+1}=0$ and calculate $x_{0,n+1}=I(\mathbf{w}'_{n+1}\boldsymbol{\gamma}+u_{0,n+1}>0)$ and then set $z_{n+1}=1$ and calculate $x_{1,n+1}=I(\mathbf{w}'_{n+1}\boldsymbol{\gamma}+u_{1,n+1}>0)$, where $u_{l,n+1}\sim\mathrm{N}(0,1/\lambda_{l,n+1})$ and $\lambda_{l,n+1}\sim$gamma $(\nu/2,\nu/2)$ $(l=0,1)$ refer to the errors given the interventions $z=0$ and 1, respectively. As prudently noted by a referee, $u_{0,n+1}$ and $u_{1,n+1}$ cannot be identical because then the potential treatments would be perfectly correlated. It should also be noted that $u_{0,n+1}$ and $u_{1,n+1}$ simply define the marginal distributions of the potential treatments; since inference about the compliance status requires just the potential treatments (not differences between the potential treatments) nothing more than these marginal distributions are needed. Furthermore, in this computation we make no assumption about the joint distribution of ($x_{0,n+1}$ and $x_{1,n+1}$) and hence no specific assumption about independence or dependence between the potential treatments.

The objective now is to calculate the distribution of $y_{j,n+1}$, truncated to the region of compliance. We denote these predictive distributions as $p_j(y_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{W},\mathbf{z},\text{Complier})$. Variates from these predictive distributions can be obtained, given each MCMC draw, as follows.

- Sample $\mathbf{w}^{(g)}_{n+1}$ by assigning probability $1/n$ to each row of $\mathbf{W}$.
- Sample $\lambda^{(g)}_{l,n+1}$ from gamma $(\nu/2,\nu/2)$, and sample $u^{(g)}_{l,n+1}$ from $\mathrm{N}(0,1/\lambda^{(g)}_{l,n+1})$ $(l=0,1)$.
- Set $z_{n+1}=0$ and calculate $x^{(g)}_{0,n+1}=I(\mathbf{w}^{(g)'}_{n+1}\boldsymbol{\gamma}^{(g)}+u^{(g)}_{0,n+1}>0)$; set $z_{n+1}=1$ and calculate $x^{(g)}_{1,n+1}=I(\mathbf{w}^{(g)'}_{n+1}\boldsymbol{\gamma}^{(g)}+\delta^{(g)}+u^{(g)}_{1,n+1}>0)$.
- Check compliance given $x^{(g)}_{0,n+1}$, $x^{(g)}_{1,n+1}$ and $\delta^{(g)}$
  - if compliant, sample $y^{(g)}_{1,n+1}$ from $\mathrm{N}(\mathbf{w}^{(g)'}_{n+1}\boldsymbol{\beta}^{(g)}_1,\sigma^{(g)2}_1/\lambda^{(g)}_{n+1})$ and $y^{(g)}_{0,n+1}$ from $\mathrm{N}(\mathbf{w}^{(g)'}_{n+1}\boldsymbol{\beta}^{(g)}_0,\sigma^{(g)2}_0/\lambda^{(g)}_{n+1})$;
  - if not compliant skip and move to the next state of the chain.

On the completion of these steps we have $\{y^{(1)}_{j,n+1},\ldots,y^{(K)}_{j,n+1}\}$ from $p_j(y_{j,n+1}|\mathbf{y},\mathbf{x},\mathbf{W},\mathbf{z},\text{Complier})$. Note that generally $K$ will be smaller than $M$ because compliance is likely to be less than perfect. We can use these generated samples to calculate the difference in means and/or the difference in quantiles.

## 5. Simulation study

### 5.1. Design

In this section we employ synthetic data to examine the properties of the fitting method and to demonstrate its viability in a high-dimensional setting with significant confounding. Realism is achieved in the simulation design by using the covariates from the work of Card (1995). In that work, Card tackles the important question of finding the effect of an additional year of schooling on a person's wage, allowing for potential confounding caused by an unobserved factor such as ability, which would likely have a positive effect on educational attainment and wages. We turn the metric schooling variable into a binary treatment variable by letting $x_i$ be an indicator of at least 12 years of schooling. The vector $\mathbf{w}_i$ is composed of the 15 covariates in that paper. The first covariate is the constant; the second and the third are experience and the square of experience; the fourth is an indicator of African-American; the fifth is an indicator of whether the subject resided in a standard metropolitan statistical area (SMSA) in 1976; the sixth is an indicator of residence in the south in 1976; the seventh is an indicator of residence in an SMSA in 1966, and the eighth to fifteenth variables are indicators to code residence in 1966 for a region variable with nine levels. The instrument is an indicator variable representing proximity to a 4-year college in 1966. The sample size $n$ is 3010.

To generate our data sets we use these covariates and the following parameter values which we round up to two decimal places:

$$\boldsymbol{\beta}_0 = (5.55, 0.09, -0.003, -0.26, 0.11, -0.26, 0.09, 0.07, 0.19, 0.11, 0.19, 0.21, 0.17, -0.05, 0.15),$$

$$\boldsymbol{\beta}_1 = (5.83, 0.08, -0.003, -0.17, 0.15, -0.10, 0.004, 0.13, 0.10, 0.01, 0.09, 0.06, 0.09, -0.06, 0.10),$$

$$(\boldsymbol{\gamma}, \delta) = (2.42, -0.43, 0.01, -0.55, 0.33, 0.22, -0.05, 0.02, 0.13, 0.23, -0.02, 0.03, -0.08, 0.76, 0.46, 0.23),$$

$$\boldsymbol{\eta}^2 = (2.00, 2.00); \quad \boldsymbol{\omega} = (0.56, 0.56)$$

for a total of 50 parameters. We arrived at the values of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$ and $(\boldsymbol{\gamma}, \delta)$ by fitting a Gaussian model to the real data. We choose the specific values of $\boldsymbol{\eta}^2$ and $\boldsymbol{\omega}$ to generate a significant amount of confounding (the correlation $\rho_j$ in each of the two treatment states is approximately 0.40). Our $i$th simulated treatment and outcome data are generated as follows.

- Simulate $\lambda_i$ from gamma $(v/2, v/2)$, where $v = 15$, and simulate $u_i \sim \mathrm{N}(0, \lambda_i^{-1})$.
- Form $x_i = I(\mathbf{w}_i'\boldsymbol{\gamma} + z_i\delta + u_i > 0)$.
- If $x_i = 1$, simulate $y_i \sim \mathrm{N}(\mathbf{w}_i'\boldsymbol{\beta}_1 + u_i\omega_1, \sigma_1^2/\lambda_i)$, where $\sigma_1^2 = \eta_1^2 - \omega_1^2$; else simulate $y_i \sim \mathrm{N}(\mathbf{w}_i'\boldsymbol{\beta}_0 + u_i\omega_0, \sigma_0^2/\lambda_i)$, where $\sigma_0^2 = \eta_0^2 - \omega_0^2$.

To get a feeling for the response variable, in our simulated data sets the average value of $y$ is between 6 and 6.5 and the proportion of treated observations is about 0.5. Finally, in our fitting, the prior distribution in Section 2.2 is parameterized by the following

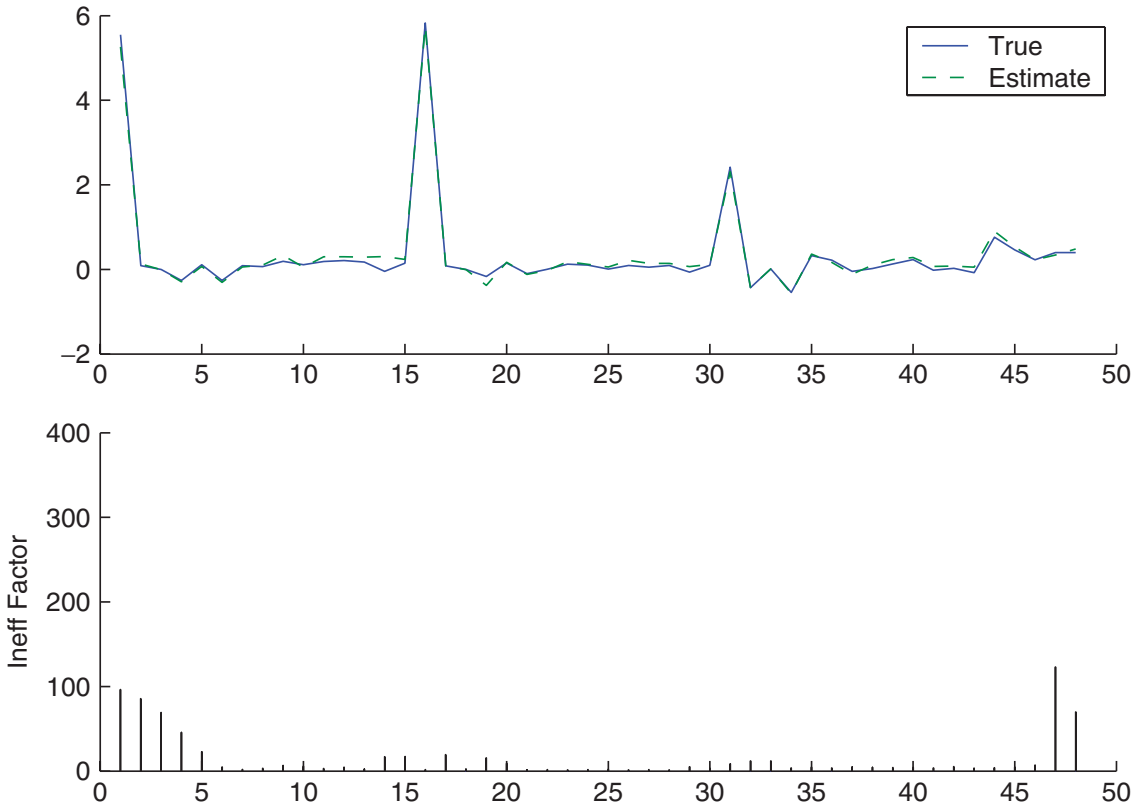


Fig. 1. Posterior mean and true values of $\boldsymbol{\beta}_0$, $\boldsymbol{\beta}_1$, $\boldsymbol{\theta}$, $\boldsymbol{\eta}$ and $\boldsymbol{\omega}$ from five simulated data sets under the student-$t$ model. MCMC sample size is 10,000, burn-in is 1000. The bottom panel gives the inefficiency factors.

hyperparameters: $n_{j0} = 4.22$; $d_{j0} = 2.22$, $j = 0, 1$ (implying a prior mean and standard deviation of 1 and 3, respectively); $m_{j,0} = 0$; $M_{j,0} = 10$; $j = 0, 1$; and $\mathbf{b}_0 = \mathbf{0}$; $\mathbf{B}_0 = 20\mathbf{I}_{46}$.

### 5.2. Results

Our results are averages of quantities calculated from five simulated data sets. A summary of the results is presented in Fig. 1 which contains the posterior means of the parameters and the corresponding true values. In the bottom panel of the figure we report the inefficiency factors (also sometimes called the autocorrelation times) which are a measure of the extent of mixing of the Markov chain output. They are obtained as one plus two times the sum of the (tapered) autocorrelations of the simulated draws. Smaller values of the inefficiency factor imply that the output is better mixing. It is clear from the figure that the estimates are close to the true values and that except for a few parameters the inefficiency factors are small.

For each of the simulated data sets we also calculate the ATE for compliers. The true value of this effect is 1.09; the average of the effect over the five fitted data sets is 1.29. We conducted additional experiments with different priors on the parameters and different true values. The results are similar to those given above and are therefore not reported.

## 6. Concluding remarks

In this paper we have shown how it is possible to develop a Bayesian framework for analyzing structural models without the joint distribution of the potential outcomes.

That this is possible has not been noticed in the literature. We present recipes for finding relevant treatment effects, averaged over both parameters and covariates, and discuss the performance of the method in simulation experiments.

Our development also makes clear that in comparison with methods that involve the unobserved counterfactuals, the approach in this paper, which is free of unnecessary counterfactuals and unidentified parameters, is easier to operationalize in more complicated problems, for example, problems with time-varying treatments, as in a panel context, and situations with an ordinal treatment. Applications of the framework discussed in this paper to these problems are ongoing and will be reported elsewhere.

## Acknowledgments

The author thanks the referees and the editor for their constructive and very helpful comments and suggestions.

## References

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

Amemiya, T., 1985. Advanced Econometrics. Harvard University Press, Boston.

Card, D., 1995. Using geographic variation in college proximity to estimate the return to schooling. In: Christophides, L.N., Grant, E.K., Swidinsky, R. (Eds.), Aspects of Labor Market Behavior. University of Toronto Press, Toronto.

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

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

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

Chib, S., Hamilton, B., 2002. Semiparametric Bayes analysis of longitudinal data treatment models. Journal of Econometrics 110, 67–89.

Dawid, A.P., 2000. Causal inference without counterfactuals (with discussion). Journal of the American Statistical Association 95, 407–424.

Dawid, A.P., 2003. Causal inference using influence diagrams: the problem of partial compliance (with discussion). In: Green, P.J., Hjort, N.L., Richardson, S. (Eds.), Highly Structured Stochastic Systems. Oxford University Press, Oxford, pp. 45–81.

Imbens, G.W., Angrist, J., 1994. Identification and estimation of local average treatment effects. Econometrica 62, 467–475.

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

Pearl, J., 2000. Causality. Cambridge University Press, Cambridge, MA.

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

Poirier, D., Tobias, J., 2003. On the predictive distributions of outcome gains in the presence of an unidentified parameter. Journal of Business and Economic Statistics 21, 258–268.

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

Vijverberg, W.P.M., 1993. Measuring the unidentified parameter of the extended Roy model of selectivity. Journal of Econometrics 57, 69–89.