Bayesian Analysis of Treatment Effects in an Ordered Potential Outcomes Model

Mingliang Li
Department of Economics
SUNY-Buffalo
mli3@buffalo.edu

Justin L. Tobias
Department of Economics
Iowa State University
tobiasj@iastate.edu

Abstract

We describe a new Bayesian estimation algorithm for fitting a binary treatment, ordered outcome selection model in a potential outcomes framework. We show how recent advances in simulation methods, namely *data augmentation*, the *Gibbs sampler* and the *Metropolis-Hastings algorithm* can be used to fit this model efficiently, and also introduce a reparameterization to help accelerate the convergence of our posterior simulator. Several computational strategies which allow for non-Normality are also discussed. Conventional "treatment effects" such as the Average Treatment Effect (ATE), the effect of treatment on the treated (TT) and the Local Average Treatment Effect (LATE) are adapted for this specific model, and Bayesian strategies for calculating these treatment effects are introduced. Finally, we review how one can potentially learn (or at least bound) the non-identified cross-regime correlation parameter and use this learning to calculate (or bound) parameters of interest beyond mean treatment effects.

ACKNOWLEDGEMENTS: We would like to thank helpful comments and suggestions from two anonymous referees, the editor Ed Vytlacil, and participants at the 4th Annual *Advances in Econometrics* Conference. All errors are our own.

1 Introduction

As evidenced by the vast literature dedicated to the issue, the problem of identifying and estimating the effects of "treatment" from observational data is of central importance to economics and the social sciences. As suggested by the articles appearing in this volume, there are many estimation strategies commonly employed in this literature, and the assumptions made in and issues emphasized by these various approaches can be quite distinct. For instance, some studies employ fully parametric models to conduct their analyses, arguing that the use of such models permits the estimation of a wide range of policy-relevant parameters,[1] while others seek a more agnostic approach and thus pursue nonparametric or semiparametric techniques.[2] Many empirical studies in this area argue that the most convincing way to surmount the problem of treatment endogeneity is to make use of cleverly chosen natural experiments or instrumental variables,[3] while others are content to pursue more structural equation approaches where the role of the exclusion restriction is decidedly less important and the discussion surrounding the instrument is muted.[4] Finally, as in econometrics generally, there are both Bayesian and Classical approaches for handling these types of models.

In this paper we focus primarily on this last distinction and take up the case of Bayesian estimation of a particular type of treatment-response model. While Bayesian work on the analysis of treatment or causal effects has become more common in the econometrics literature [e.g., Vijverberg (1993), Koop and Poirier (1997), Li (1998), Chib and Hamilton (2000, 2002), Poirier and Tobias (2003) and Li, Poirier and Tobias (2004)], the use of such techniques continues to remain rare relative to Classical approaches. We do not aim to reduce this disparity by proselytizing at length in this paper about the merits of the Bayesian approach relative to Classical methods. Instead, our goal is to review how a Bayesian might handle specifications, similar to the Roy (1951) model, which are commonly encountered in the treatment effect literature, to review some computational advances which should appeal

[1] Heckman, Tobias and Vytlacil (2003), for example, discuss parametric approaches for estimating a variety of popular treatment effects under various distributional assumptions.

[2] Manski's (1990, 1994) nonparametric bounding is a leading example.

[3] See Angrist and Krueger (2001) for a review.

[4] Gould (2002,2005), for example, argues that having strong predictors for treatment status is more important for practical identification purposes than requiring that some set of covariates are excluded from the outcome equation. In applied Bayesian work [e.g., Poirier and Tobias (2003), Munkin and Trivedi (2003) and Li, Poirier and Tobias (2004)], the instrument tends to receive decidedly less discussion. In empirical practice, however, such exclusion restrictions should be, and typically are used when available.

to all researchers when faced with estimation of these types of models, to introduce an issue that is somewhat unique to the Bayesian literature on this topic, and to provide new results on Bayesian estimation of a specific type of treatment effect model.

We take up the particular case of a treatment-response model where treatment status is binary and the outcome of interest is ordered. To our knowledge, a discussion of this particular model is new to the Bayesian literature, though highly related models, including those of the binary treatment / continuous outcome and ordered treatment / binary outcome varieties have appeared in Chib and Hamilton (2000). We present our model in a *potential outcomes* framework and thus model both the observed outcome of the agent given her treatment choice as well as the potential or counterfactual outcome for that agent had she made a different treatment decision.

We show how *data augmentation* [e.g., Tanner and Wong (1987), Albert and Chib (1993)] in conjunction with the *Gibbs sampler* and *Metropolis-Hastings algorithm* [e.g., Casella and George (1992), Tierney (1994), Chib and Greenberg (1995)] can be used to fit this particular model efficiently, and also introduce a reparameterization to help accelerate the convergence of our posterior simulator. Several computational strategies which allow for non-Normality are also discussed, though not employed. Treatment effects similar in spirit to the Average Treatment Effect (ATE), the effect of treatment on the treated (TT) and the Local Average Treatment Effect (LATE)[5] are adapted for the case of our ordered response, and Bayesian strategies for calculating these treatment effects are described. Finally, we discuss how one can potentially learn about (or at least bound) the non-identified cross-regime correlation parameter[6] and use this learning to calculate (or bound) parameters of interest beyond mean treatment effects.

The outline of this paper is as follows. Section 2 presents the basic potential outcomes model and section 3 discusses our Bayesian estimation algorithm. Often-reported treatment parameters such as ATE, TT and LATE are derived for our model in section 4 and procedures for calculating these effects are described. A generated data experiment which illustrates the performance of our algorithm is provided in section 5, and the paper concludes with a

[5] See, for example, Imbens and Angrist (1994) for a discussion of LATE and Heckman and Vytlacil (1999, 2000) for detailed discussions of these and other treatment effects.

[6] For related discussions on this topic, see Vijverberg (1993), Koop and Poirier (1997), Poirier (1998), Poirier and Tobias (2003) and Li, Poirier and Tobias (2004).

summary in section 6.

2 The Model

What we have in mind is the development of a parametric model that will enable researchers to investigate the impact of a binary (and potentially endogenous) treatment variable, denoted D, where $D = 1$ implies receipt of treatment and $D = 0$ implies non-receipt, on an *ordered* outcome of interest, denoted $y \in \{1, 2, \cdots, J\}$. There are numerous examples where such a model would be appropriate. For example, one might use this model to investigate, say, the impact of enrolling in a supplemental learning center on attitudes toward education (measured as a categorical response) or the quantity of education ultimately received by the student. More generally, such a model is potentially of value in any situation where the outcome of interest (e.g., earnings, education, expenditure) is recorded categorically rather than continuously and the model also contains a dummy endogenous variable.[7]

We cast this evaluation problem in a *potential outcomes* framework and thus explicitly model the *counterfactual* state - the ordered outcome that would have been observed had the agent made a different treatment decision. We let $y^{(1)}$ denote the outcome received by the agent in the treatment state and $y^{(0)}$ denote the outcome received without treatment. Only one outcome, denoted y_i, is ever observed for any agent, and thus

$$y_i = D_i y_i^{(1)} + (1 - D_i) y_i^{(0)}.$$

We suppose that the observed treatment decision D and the observed and potential ordered outcomes $y^{(1)}$ and $y^{(0)}$ are generated by an underlying latent variable representation of the model. Specifically, we write:[8]

$$\begin{aligned} D_i^* &= w_i \beta^{(D)} + u_i & (1) \\ z_i^{(1)} &= x_i \beta^{(1)} + \epsilon_i^{(1)} & (2) \end{aligned}$$

[7] One can also conceive of situations where the modeling of count outcomes is desired (*e.g.* Munkin and Trivedi 2003). Clearly, approaches to modeling ordered and count outcomes impose different parametric assumptions on the response (*e.g.* ordered probit versus Poisson or negative binomial distribution), invoke different interpretations of the outcomes of interest (ordinal versus cardinal), and involve different assessments of the censoring feature of the outcomes (censored versus unbounded). Which approach is more appropriate depends critically on the type of application that is considered.

[8] In this paper, we assume that the same set of covariates appear in the treated and untreated states. If desired, this assumption could be relaxed and this extension incorporated into the derivations which follow.

$$z_i^{(0)} \quad = \quad x_i\beta^{(0)} + \epsilon_i^{(0)}. \tag{3}$$

The binary treatment indicator D_i is related to the latent D_i^* as follows:

$$D_i = I(D_i^* > 0) = I[u_i > -w_i\beta^{(D)}], \tag{4}$$

with $I(\cdot)$ denoting the standard indicator function. Similarly, the ordered responses $y_i^{(1)}$ and $y_i^{(0)}$ are related to the latent variables $z_i^{(1)}$ and $z_i^{(0)}$ as follows:

$$y_i^{(k)} = j \text{ iff } \alpha_j^{(k)} < z_i^{(k)} \leq \alpha_{j+1}^{(k)}, \quad k = 0, 1, \quad j = 1, 2, \cdots, J. \tag{5}$$

The $\{\alpha_j^{(k)}\}, k = 0, 1, \ j = 1, 2, \cdots, J$ are *cutpoints* in the model, mapping the latent indices in both states into discrete values of our ordered response. We impose standard identification conditions on these cutpoints, namely, $\alpha_1^{(1)} = \alpha_1^{(0)} = -\infty$, $\alpha_2^{(1)} = \alpha_2^{(0)} = 0$ and $\alpha_{J+1}^{(1)} = \alpha_{J+1}^{(0)} = \infty$. We also let

$$\alpha^{(1)} = [\alpha_3^{(1)} \ \alpha_4^{(1)} \ \cdots \ \alpha_J^{(1)}]$$

denote the cutpoint vector for the treated state and define the $1 \times (J-2)$ vector $\alpha^{(0)}$ similarly.

In this model we also assume the availability of an *exclusion restriction* - some covariate which enters w that is not contained in x. To motivate the importance of this assumption, consider a restricted version of (1)-(3) which consists of equation (1) and a latent variable equation like (2), the latter of which includes the observed D_i as an element of x_i. This restricted model would be of the form of a "standard" treatment or causal effect model that only works with observed rather than potential outcomes. Maddala (1983, page 122), for example, shows that the parameters of such a model are not identifiable unless the errors of the equation system are uncorrelated or such an exclusion restriction is present. The former condition often seems rather untenable in empirical practice, and thus we maintain that such an exclusion restriction is available.

Finally, we fix ideas throughout the remainder of this discussion by assuming joint Normality of the error terms:[9]

$$\begin{bmatrix} u_i \\ \epsilon_i^{(1)} \\ \epsilon_i^{(0)} \end{bmatrix} | x_i, w_i \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \rho^{(1)} & \rho^{(0)} \\ \rho^{(1)} & 1 & \rho^{(10)} \\ \rho^{(0)} & \rho^{(10)} & 1 \end{bmatrix} \right) \equiv N(0, \Sigma). \tag{6}$$

Equations (1) - (6) then denote the complete specification of our ordered potential outcomes model.

[9] We discuss how this requirement can be relaxed in section 3.4 of this paper. The variances of the errors in all the equations have been normalized to unity for identification purposes.

2.1 The likelihood

Given the assumed conditional independence across observations, we can write the likelihood function for this model as:

$$p(y, D|\Gamma) \equiv L(\Gamma; y, D) = [\prod_{i:D_i=1} \Pr(y_i^{(1)} = y_i, D_i = 1|\Gamma)][\prod_{i:D_i=0} \Pr(y_i^{(0)} = y_i, D_i = 0|\Gamma)],$$

where $\Gamma = [\beta^{(D)}\ \beta^{(1)}\ \beta^{(0)}\ \alpha^{(1)}\ \alpha^{(0)}\ \rho^{(0)}\ \rho^{(1)}\ \rho^{(10)}]$. The joint probabilities required in calculating this likelihood can be obtained from the bivariate Normal cdf. For example,

$$\begin{aligned} \Pr(y_i^{(1)} = y_i, D_i = 1|\Gamma) &= \Pr(\alpha_{y_i}^{(1)} < z_i^{(1)} \leq \alpha_{y_i+1}^{(1)},\ u_i > -w_i\beta^{(D)}|w_i, x_i, \Gamma) && (7) \\ &= \Pr(\alpha_{y_i}^{(1)} - x_i\beta^{(1)} < \epsilon_i^{(1)} \leq \alpha_{y_i+1}^{(1)} - x_i\beta^{(1)},\ u_i > -w_i\beta^{(D)}|w_i, x_i, \Gamma). && (8) \end{aligned}$$

Provided one uses a statistical package containing a routine for evaluating a bivariate Normal cdf, standard MLE can be implemented. If no such routine is available on a particular package, one could first reduce the probabilities above to univariate integration problems and then employ standard numerical approximations such as Simpson's rule or Gaussian quadrature to approximate the requisite integrals. To see this more clearly, let $P_{1,j} \equiv \Pr(D_i = 1, y_i^{(1)} = j|\Gamma)$, and note from (8)

$$\begin{aligned} P_{1,j} &= \Pr(\alpha_j^{(1)} - x_i\beta^{(1)} < \epsilon_i^{(1)} \leq \alpha_{j+1}^{(1)} - x_i\beta^{(1)},\ u_i > -w_i\beta^{(D)}|x_i, w_i, \Gamma) \\ &= \int_{\alpha_j^{(1)} - x_i\beta^{(1)}}^{\alpha_{j+1}^{(1)} - x_i\beta^{(1)}} \int_{-w_i\beta^{(D)}}^{\infty} p(\epsilon_i^{(1)}, u_i)\ du_i\ d\epsilon_i^{(1)} \\ &= \int_{\alpha_j^{(1)} - x_i\beta^{(1)}}^{\alpha_{j+1}^{(1)} - x_i\beta^{(1)}} \Phi\left(\frac{w_i\beta^{(D)} - \rho^{(1)}\epsilon_i^{(1)}}{\sqrt{1 - \rho^{(1)^2}}}\right) p(\epsilon_i^{(1)})\ d\epsilon_i^{(1)}. \end{aligned}$$

In this form a variety of approaches can be employed to approximate the required univariate integrals. In our discussion of treatment effects in section 4, we will return to one approach to this problem based on Monte Carlo integration using truncated Normal sampling. Importantly, we also recognize that our estimation strategy via data augmentation, as described in the following section, avoids the need for any numerical integration of the above form, and therefore provides an attractive alternative to the implementation of standard MLE.

3 Bayesian Estimation

To perform a Bayesian analysis, a researcher first starts off as a classical econometrician might by specifying the likelihood function for this model, as implied from (1) - (6) and described in the preceding section. To this likelihood, the researcher adds a *prior density*, say $p(\Gamma)$, with Γ denoting the parameters of the model. This prior is chosen to reflect her subjective beliefs about values of the parameters, and in most cases is chosen to be sufficiently "vague" or "flat" so that information contained in the data will dominate information insinuated through the prior. The prior density $p(\Gamma)$ combined with the likelihood $p(y, D|\Gamma)$ yields the joint posterior density $p(\Gamma|y, D)$ up to proportionality via Bayes theorem. This joint posterior completely summarizes the "output" of a Bayesian procedure - from it, one can obtain point and interval estimates, marginal posterior densities, posterior quantiles, or other quantities of interest.

While in theory this simple exercise outlines *the* machinery involved in Bayesian posterior calculations, in practice, extracting useful information from a given posterior $p(\Gamma|y, D)$ can be difficult. Direct calculation of a posterior mean of an element of Γ, for example, first requires that the normalizing constant of the joint posterior is known (while often it is not), and even if the normalizing constant were known, the mean calculation would still require solving a high-dimensional integration problem. In models of moderate complexity, these integration problems usually have no analytic solutions.

Instead of direct evaluation of this posterior, modern Bayesian empirical work makes use of recent advances in simulation methods to carry out a posterior analysis. Two simulation devices in particular, called the *Gibbs sampler* and *Metropolis-Hastings algorithm*, are widely used and have become indispensable instruments in an applied Bayesian's toolkit. Both of these algorithms solve the problem of calculation of posterior moments, quantiles, marginal densities or other quantities of interest by first obtaining a set of draws from the posterior $p(\Gamma|y, D)$. Typically, one can not draw directly from this density, but instead, one can generate a sequence of draws (by appropriately following the steps of the algorithms) that converge to this distribution. Once convergence has been "achieved," the subsequent set of simulated parameter values can be used to calculate the desired quantities (e.g., posterior means). In the Gibbs sampler, a Markov Chain whose limiting distribution is $p(\Gamma|y, D)$ is produced by iteratively sampling form the complete posterior conditionals of the model. In many cases, typically in models with conditionally conjugate priors, these posterior conditionals have well-known forms and can be easily sampled. The Metropolis-Hastings algorithm is a generalization of the Gibbs sampler and is a multivariate accept-reject algorithm. The algorithm is, again, constructed so that the limiting distribution of the Markov chain is the target density, $p(\Gamma|y, D)$.[10]

In terms of the model described in this paper, Bayesian estimation of the specification in (1) - (6) would likely make use of *data augmentation* [e.g. Tanner and Wong (1987), Albert and Chib (1993)] in conjunction with the algorithms above. When data augmentation is used, the posterior is first expanded (or, as the name suggests, *augmented*) to include not only the parameter vector Γ, but also the latent data $s = [D^* \; z^{(1)} \; z^{(0)}]$. Although this would seem to complicate the estimation exercise, use of data augmentation often simplifies the required posterior calculations. This is particularly true when data augmentation is used in conjunction with the Gibbs sampler since, conditioned on the latent data, inference regarding the regression parameters proceeds as a linear regression model would, and given the regression parameters, it is often straight-forward to obtain draws from the posterior conditional for the latent data.

For our model, this *augmented* posterior is of the form:

$$\begin{aligned} p(D^*, z^{(1)}, z^{(0)}, \Gamma|y, D) &\propto p(y, D, D^*, z^{(1)}, z^{(0)}, \Gamma) & (9) \\ &= p(y, D|D^*, z^{(1)}, z^{(0)}, \Gamma)p(D^*, z^{(1)}, z^{(0)}|\Gamma)p(\Gamma), & (10) \end{aligned}$$

with $p(\Gamma)$ denoting the prior for the parameters of our model. The middle term in the above expression is immediately known as a trivariate Normal density, given the joint Normality assumption in (6) combined with the model in (1)-(3). The last term simply denotes the prior for our model parameters. For the first term, conditioned on the latent variables and model parameters, the observed responses D and y are known with certainty and thus the joint (conditional) distribution for y and D is degenerate. Putting these pieces together, and exploiting the assumed conditional independence across observations, we can write the

[10] A detailed review of these simulation methods is beyond the scope of this paper; the interested reader is invited to see Casella and George (1992), Tierney (1994), Chib and Greenberg (1995), Gilks et al (1998), Geweke (1999), Chen, Shao and Ibrahim (2000), Carlin and Louis (2000), Geweke and Keane (2001), Chib (2001), Koop (2003), Lancaster (2004), Gelman et al (2004), Poirier and Tobias (2006) and Koop, Poirier and Tobias (2006) (among others) for detailed and comprehensive descriptions of these and other methods.

augmented posterior as follows:

$$p(D^*, z^{(1)}, z^{(0)}, \Gamma | D, y) \propto p(\Gamma) \prod_{i=1}^{n} \phi_3(s_i; r_i\beta, \Sigma) \times \tag{11}$$
$$\Big[I(D_i = 1)I(D_i^* > 0)I(\alpha_{y_i}^{(1)} < z_i^{(1)} \leq \alpha_{y_i+1}^{(1)}) + I(D_i = 0)I(D_i^* \leq 0)I(\alpha_{y_i}^{(0)} < z_i^{(0)} \leq \alpha_{y_i+1}^{(0)}) \Big].$$

In the above, we have defined

$$s_i = \begin{bmatrix} D_i^* \\ z_i^{(1)} \\ z_i^{(0)} \end{bmatrix}, \; r_i = \begin{bmatrix} w_i & 0 & 0 \\ 0 & x_i & 0 \\ 0 & 0 & x_i \end{bmatrix}, \; \beta = \begin{bmatrix} \beta^{(D)} \\ \beta^{(1)} \\ \beta^{(0)} \end{bmatrix}, \tag{12}$$

and $\phi_3(x; \mu, \Omega)$ denotes a trivariate Normal density with mean μ and covariance matrix Ω. Finally, Σ is defined in (6). The indicator functions added to (11) serve to capture the degenerate joint distribution of y and D given the latent data and model parameters.

3.1 A Useful Reparameterization

In theory, one could directly apply standard computational tools (namely the Gibbs Sampler coupled with a few Metropolis-within-Gibbs steps) to fit the model in (11). However, it has been shown in related work [e.g., Cowles (1996), Nandram and Chen (1996) and Li and Tobias (2005)], that use of the standard Gibbs sampler in models with ordered responses suffers from slow mixing due to high correlation between the simulated cutpoints and latent data. As discussed in the previous section, the parameter draws obtained from our estimation algorithm form a Markov chain, and when the chain mixes slowly, we observe only very small local movements from iteration to iteration. As a result, it may take a very long time for our simulator to traverse the entire parameter space. When the lagged autocorrelations between the simulated parameters are very high, estimates of posterior features may be quite inaccurate, and numerical standard errors associated with those estimates will be unacceptably large. To mitigate this slow mixing problem, and move closer to a situation where we can obtain iid samples from the posterior, we suggest below an alternate parameterization of the model, building of the suggestion of Nandram and Chen (1996).

To shed some insight on this reparameterization, first separate out the largest cutpoints from the treated state, $(\alpha_J^{(1)})$, and untreated state, $(\alpha_J^{(0)})$, and define the transformations:

$$\sigma_1 = 1/[\alpha_J^{(1)}]^2 \quad \text{and} \quad \sigma_0 = 1/[\alpha_J^{(0)}]^2.$$

In addition, for any variable Q_i let $\tilde{Q}_i^{(1)} \equiv \sqrt{\sigma_1} Q_i^{(1)}$ and define $\tilde{Q}_i^{(0)} \equiv \sqrt{\sigma_0} Q_i^{(0)}$ similarly.

The model in (1) - (3) is then *observationally equivalent* to

$$\begin{aligned} D_i^* &= w_i \beta^{(D)} + u_i && (13) \\ \tilde{z}_i^{(1)} &= x_i \tilde{\beta}^{(1)} + \tilde{\epsilon}_i^{(1)} && (14) \\ \tilde{z}_i^{(0)} &= x_i \tilde{\beta}^{(0)} + \tilde{\epsilon}_i^{(0)}. && (15) \end{aligned}$$

where

$$y_i^{(k)} = j \quad \text{iff} \quad \tilde{\alpha}_j^{(k)} < \tilde{z}_i^{(k)} \leq \tilde{\alpha}_{j+1}^{(k)}, \quad k = 0, 1. \tag{16}$$

In other words, the likelihood function for the observed data is unchanged when multiplying (2) and (3) by $\sqrt{\sigma_1}$ and $\sqrt{\sigma_0}$, respectively, and appropriately adjusting the rule in (16) which maps the latent data into the observed responses. The error variance for the transformed disturbances now takes the following form:

$$\begin{bmatrix} u_i \\ \tilde{\epsilon}_i^{(1)} \\ \tilde{\epsilon}_i^{(0)} \end{bmatrix} | x_i, w_i \overset{iid}{\sim} N\left(\begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & \tilde{\sigma}_{1D} & \tilde{\sigma}_{0D} \\ \tilde{\sigma}_{1D} & \sigma_1 & \tilde{\sigma}_{10} \\ \tilde{\sigma}_{0D} & \tilde{\sigma}_{10} & \sigma_0 \end{bmatrix} \right) \equiv N(0, \tilde{\Sigma}), \tag{17}$$

where $\tilde{\sigma}_{1D} \equiv \sqrt{\sigma_1} \rho^{(1)}$, $\tilde{\sigma}_{0D} \equiv \sqrt{\sigma_0} \rho^{(0)}$ and $\tilde{\sigma}_{10} = \sqrt{\sigma_1}\sqrt{\sigma_0} \rho^{(10)}$. The correlation parameters $\rho^{(1)}, \rho^{(0)}$ and $\rho^{(10)}$ are defined in (6).

When the model is written as in (13) - (17), it suggests that we can work with an augmented posterior distribution containing the latent variables $D^*, \tilde{z}^{(1)}, \tilde{z}^{(0)}$ and parameters $\tilde{\Gamma} = [\tilde{\beta} \ \tilde{\sigma}_{1D} \ \tilde{\sigma}_{0D} \ \tilde{\sigma}_{10} \ \sigma_1 \ \sigma_0 \ \tilde{\alpha}^{(1)} \ \tilde{\alpha}^{(0)}]$ instead of D^*, $z^{(1)}$, $z^{(0)}$ and Γ as in (11). The transformed cutpoint and coefficient vectors contained in $\tilde{\Gamma}$ are defined as follows:[11]

$$\tilde{\alpha}^{(1)} = [\tilde{\alpha}_3^{(1)} \ \tilde{\alpha}_4^{(1)} \ \ \cdots \tilde{\alpha}_{J-1}^{(1)}], \quad \tilde{\alpha}^{(0)} = [\tilde{\alpha}_3^{(0)} \ \tilde{\alpha}_4^{(0)} \ \ \cdots \tilde{\alpha}_{J-1}^{(0)}], \quad \text{and} \quad \tilde{\beta} = [\beta^{(D)} \ \tilde{\beta}^{(1)} \ \tilde{\beta}^{(0)}].$$

Following similar derivations to those leading to (11), we obtain the *augmented* joint posterior distribution for the transformed parameters:

$$\begin{aligned} & p(D^*, \tilde{z}^{(1)}, \tilde{z}^{(0)}, \tilde{\Gamma} | D, y) \propto p(\tilde{\Gamma}) \prod_{i=1}^{n} \phi_3(\tilde{s}_i; r_i \tilde{\beta}, \tilde{\Sigma}) \times && (18) \\ & \left[I(D_i = 1) I(D_i^* > 0) I(\tilde{\alpha}_{y_i}^{(1)} < \tilde{z}_i^{(1)} \leq \tilde{\alpha}_{y_i+1}^{(1)}) + I(D_i = 0) I(D_i^* \leq 0) I(\tilde{\alpha}_{y_i}^{(0)} < \tilde{z}_i^{(0)} \leq \tilde{\alpha}_{y_i+1}^{(0)}) \right] \end{aligned}$$

where $\tilde{s}_i \equiv [D_i^* \ \tilde{z}_i^{(1)} \ \tilde{z}_i^{(0)}]'$ and $\tilde{\Sigma}$ is defined in (17).

[11] Note that the largest cutpoints have been taken out of each cutpoint vector and these largest cutpoints are replaced by σ_1 and σ_0 in this alternate parameterization.

When working with this model, we employ independent priors for the parameters of $\tilde{\Gamma}$:

$$p(\tilde{\Gamma}) = p(\tilde{\beta})p(\tilde{\alpha}^{(1)})p(\tilde{\alpha}^{(0)})p(\tilde{\Sigma}).$$

We center the regression parameters around a prior mean of zero and specify them to be independently distributed with large prior variances: $\tilde{\beta} \sim N(b_0 = 0_{k\times 1}, V_{\tilde{\beta}} = 1000 I_k)$. The prior probability density function of $\tilde{\alpha}^{(1)}$ and $\tilde{\alpha}^{(0)}$ is assumed to be proportional to some constant: $p(\tilde{\alpha}_3^{(1)}, \cdots, \tilde{\alpha}_{J-1}^{(1)}, \tilde{\alpha}_3^{(0)}, \cdots, \tilde{\alpha}_{J-1}^{(0)}) \propto c$, and finally, an inverse Wishart prior of the form $\tilde{\Sigma} \sim IW(\underline{\rho}, \underline{R})$ with $\underline{\rho} = 6$, $\underline{R} = I_3$ is employed subject to the restriction that the $(1,\ 1)$ element of $\tilde{\Sigma}$ is equal to one.

3.2 Benefits and Costs of Reparameterization

To this point we have offered no compelling arguments why one should work with the reparameterized model instead of working directly with the original "structural" representation of the model. The first argument in support of the reparameterization, as noted in Nandram and Chen (1996), and further shown in Li and Tobias (2005), is that the rescaling helps to significantly reduce the autocorrelation among the posterior simulations, thus accelerating the convergence of the algorithm. In other words, given an equal number of posterior draws, the numerical standard errors obtained when working with the reparameterized model will be significantly smaller than those obtained when using the original parameterization of the model. Second, and quite importantly from a computational point of view, this transformation effectively "restores" the conjugacy required to simulate the parameters of the covariance matrix. That is, in the original parameterization of the model in (6), there are restrictions on all three diagonal elements of the covariance matrix. This precludes drawing the elements of the inverse covariance matrix from a Wishart distribution, (as is typically the case when conjugate priors are employed), since the posterior conditional is no longer Wishart given the diagonal restrictions. On the other hand, in our reparameterized version of the model, the covariance matrix in (17) contains only one diagonal restriction, and using Algorithm 3 of Nobile (2000), one can generate draws from this restricted Wishart density. Thus, working with the reparameterized model facilitates simulating parameters of the covariance matrix, and no Metropolis-Hastings steps are required for this portion of the posterior simulator. Finally, for the specific case where there are three possible ordered outcomes (i.e., $J = 3$), there are effectively no unknown cutpoints in the transformed representation of this model.

For this case, the cutpoints are sampled through standard sampling of the elements of the covariance matrix. For this particular model with 3 outcomes, posterior simulation using the reparameterized model is quite fast, and no Metropolis-Hastings steps are required at *any* point in the algorithm. For $J > 3$, however, additional Metropolis-Hastings steps are required to simulate elements of the cutpoint vectors $\tilde{\alpha}_1$ and $\tilde{\alpha}_0$.

The main, and perhaps only, drawback to working with the reparameterized model is that it requires us to place priors on the transformed parameters $\tilde{\Gamma}$. The priors we place on these parameters may seem reasonable and suitably "default," but upon closer investigation, they may imply priors for the structural parameters that are unreasonable and are at odds with our views about quantities for which we can more easily elicit our prior beliefs. For a two-equation treatment-response model containing an ordered treatment variable and an ordered response, Li and Tobias (2005) derive some connections between priors like those employed in (18) and their consequences on the priors implied on the structural parameters in (11). With suitably chosen hyperparameters, they argue that the implied priors on the structural coefficients can be reasonable, and that any costs associated with this prior selection issue are more than outweighed by the benefits afforded by the reparameterization. We take up a more detailed view of this issue of prior selection for this particular model in our generated data experiments of section 5.

3.3 The Posterior Simulator

We now introduce our posterior simulator for fitting our reparameterized treatment-response model. In what follows, we adopt the notation $\tilde{\Gamma}_{-x}$ to denote all parameters other than x. We first group the joint posterior into $[D^* \quad \tilde{z}^{(1)} \ \tilde{z}^{(0)} \ \tilde{\alpha}^{(1)} \ \tilde{\alpha}^{(0)} \ \tilde{\beta} \ \tilde{\Sigma}]$. The latent data and cutpoints will be sampled in *blocking steps*, while the regression parameters and covariance matrix will be drawn from their complete posterior conditional.

Step 1: Draw $\tilde{\beta}$ from[12]

$$\tilde{\beta}|\tilde{s}, \tilde{\Gamma}_{-\tilde{\beta}}, y, D \sim N\left(D_{\tilde{\beta}} d_{\tilde{\beta}}, D_{\tilde{\beta}}\right)$$

[12] It is useful to note that, conditional on the latent variables, our model is essentially a seemingly unrelated regressions (SUR) model except for the restriction that one diagonal element of the covariance matrix is fixed at one.

where

$$D_{\tilde{\beta}} \equiv \left(\sum_{i=1}^{n} r_i' \tilde{\Sigma}^{-1} r_i + V_{\tilde{\beta}}^{-1} \right)^{-1} \text{ and } d_{\tilde{\beta}} \equiv \sum_{i=1}^{n} r_i' \tilde{\Sigma}^{-1} \tilde{s}_i + V_{\tilde{\beta}}^{-1} b_0.$$

Step 2: Draw $\tilde{\Sigma}$ from

$$\tilde{\Sigma} | \tilde{s}, \tilde{\Gamma}_{-\tilde{\Sigma}}, y, D \sim IW \left(n + \underline{\rho}, \left[\sum_{i=1}^{n} (\tilde{s}_i - r_i \tilde{\beta})(\tilde{s}_i - r_i \tilde{\beta})' + \underline{\rho} \underline{R} \right] \right) I(\tilde{\Sigma}_{11} = 1).$$

Algorithm 3 in Nobile (2000) is used to generate variates from this inverted Wishart distribution, conditioned on the value of the (1,1) element.

The remaining steps in the posterior simulator involve joint sampling of the latent data $\tilde{s} = [D^* \; \tilde{z}^{(1)} \; \tilde{z}^{(0)}]$ and cutpoint vectors $\tilde{\alpha}^{(1)}$ and $\tilde{\alpha}^{(0)}$. We attempt to mitigate autocorrelation in our parameter chains by *blocking* or *grouping* the cutpoints from a given equation together with the latent data appearing in that equation. Specifically, we proceed by sampling from the following densities:

$$\tilde{\alpha}^{(1)}, \tilde{z}^{(1)} | \tilde{z}^{(0)}, D^*, \tilde{\Gamma}_{-\tilde{\alpha}^{(1)}}, y, D \tag{19}$$

$$\tilde{\alpha}^{(0)}, \tilde{z}^{(0)} | \tilde{z}^{(1)}, D^*, \tilde{\Gamma}_{-\tilde{\alpha}^{(0)}}, y, D \tag{20}$$

and

$$D^* | \tilde{z}^{(1)}, \tilde{z}^{(0)}, \tilde{\Gamma}, y, D. \tag{21}$$

Taking a closer look at the first of these three densities, we find from (18)

$$\begin{aligned} \tilde{\alpha}^{(1)}, \tilde{z}^{(1)} | \tilde{z}^{(0)}, D^*, \tilde{\Gamma}_{-\tilde{\alpha}^{(1)}}, y, D \quad & \propto \\ \prod_{i=1}^{n} \phi_3(\tilde{s}_i; r_i \tilde{\beta}, \tilde{\Sigma}) & [I(\tilde{\alpha}_{y_i}^{(1)} < \tilde{z}_i^{(1)} \leq \tilde{\alpha}_{y_i+1}^{(1)}) I(D_i = 1) + I(D_i = 0)] \\ \propto \prod_{i=1}^{n} \phi(\tilde{z}_i^{(1)}; \tilde{\mu}_{i1}^c, \tilde{\sigma}_1^c) & [I(\tilde{\alpha}_{y_i}^{(1)} < \tilde{z}_i^{(1)} \leq \tilde{\alpha}_{y_i+1}^{(1)}) I(D_i = 1) + I(D_i = 0)]. \end{aligned} \tag{22}$$

Note that the indicator functions involving D_i^* and $\tilde{z}_i^{(0)}$ in (18) have disappeared completely simply because we are now conditioning on these latent parameters. In the last line of (22), we have broken the trivariate Normal density for $\tilde{s}_i$ into a conditional for $\tilde{z}_i^{(1)}$ times the joint for $\tilde{z}_i^{(0)}$ and D_i^*. The latter joint density is then absorbed into the normalizing constant of (22), as it does not involve $\tilde{\alpha}^{(1)}$ or $\tilde{z}^{(1)}$. It follows that the conditional mean $\tilde{\mu}_{i1}^c$ and conditional variance $\tilde{\sigma}_1^c$ are defined as:

$$\tilde{\mu}_{i1}^c \equiv x_i \tilde{\beta}^{(1)} + [\tilde{\sigma}_{1D} \; \tilde{\sigma}_{10}] \begin{bmatrix} 1 & \tilde{\sigma}_{0D} \\ \tilde{\sigma}_{0D} & \sigma_0 \end{bmatrix}^{-1} \begin{bmatrix} D_i^* - w_i \beta^{(D)} \\ \tilde{z}_i^{(0)} - x_i \beta^{(0)} \end{bmatrix}$$

and

$$\tilde{\sigma}_1^c \equiv \sigma_1 - [\tilde{\sigma}_{1D}\ \tilde{\sigma}_{10}] \begin{bmatrix} 1 & \tilde{\sigma}_{0D} \\ \tilde{\sigma}_{0D} & \sigma_0 \end{bmatrix}^{-1} \begin{bmatrix} \tilde{\sigma}_{1D} \\ \tilde{\sigma}_{10} \end{bmatrix}.$$

To obtain a draw from (19), we proceed in two steps and use the *method of composition* [see, e.g., Chib (2001)]. First, we marginalize (19) over $\tilde{z}^{(1)}$ and describe a procedure for drawing $\tilde{\alpha}^{(1)}$ from this density. In the second step, we draw $\tilde{z}^{(1)}$ from $\tilde{z}^{(1)}|\tilde{z}^{(0)}, D^*, \tilde{\Gamma}, y, D$. The realized values of $\tilde{\alpha}^{(1)}$ and $\tilde{z}^{(1)}$ then form a draw from (19).

After integrating (19) over $\tilde{z}^{(1)}$, we obtain:

$$\tilde{\alpha}^{(1)}|\tilde{z}^{(0)}, D^*, \tilde{\Gamma}_{-\tilde{\alpha}^{(1)}}, y, D \propto \prod_{i:D_i=1} \Phi\left(\frac{\tilde{\alpha}^{(1)}_{y_i+1} - \tilde{\mu}^c_{i1}}{\sqrt{\tilde{\sigma}^c_1}}\right) - \Phi\left(\frac{\tilde{\alpha}^{(1)}_{y_i} - \tilde{\mu}^c_{i1}}{\sqrt{\tilde{\sigma}^c_1}}\right). \tag{23}$$

Step 3: To sample from the density in (23), we follow the suggestion of Nandram and Chen (1996), who suggest using a Dirichlet proposal density to sample *differences*[13] between the cutpoint values, $q_j^{(1)} = \tilde{\alpha}^{(1)}_{j+1} - \tilde{\alpha}^{(1)}_j, \quad j = 3, \cdots, J-1$. Given that the largest cutpoint takes the value of unity, we can then solve back to obtain the values of the cutpoints themselves. Specifically, we sample $\{q_j^{(1)}\}_{j=3}^{J-1} \sim Dirichlet(\{\delta_j^{(1)} n_j^{(1)} + 1\}_{j=3}^{J-1})$, where $\{\delta_j^{(1)}\}_{j=3}^{J-1} = 0.1$ are tuning parameters, and $n_j^{(1)} \equiv \sum_{i=1}^n I(y_i = j) I(D_i = 1),\ j = 3, \cdots J-1$ are the numbers of individuals falling into each category of the outcome variable in the treated state. The probability of accepting the candidate draw is the standard Metropolis-Hastings probability, $\min(R, 1)$, where

$$R = [\prod_{i:D_i=1} \frac{\Phi([\tilde{\alpha}^{(1)}_{y_i+1,can} - \tilde{\mu}^c_{i1}]/\sqrt{\tilde{\sigma}^c_1}) - \Phi([\tilde{\alpha}^{(1)}_{y_i,can} - \tilde{\mu}^c_{i1}]/\sqrt{\tilde{\sigma}^c_1})}{\Phi([\tilde{\alpha}^{(1)}_{y_i+1,l-1} - \tilde{\mu}^c_{i1}]/\sqrt{\tilde{\sigma}^c_1}) - \Phi([\tilde{\alpha}^{(1)}_{y_i,l-1} - \tilde{\mu}^c_{i1}]/\sqrt{\tilde{\sigma}^c_1})}][\prod_{j=3}^{J-1} (\frac{q^{(1)}_{j,l-1}}{q^{(1)}_{j,can}})^{\delta_j^{(1)} n_j^{(1)}}],$$

"$_{l-1}$" denotes the current value of the algorithm and "can" denotes the candidate draw from the Dirichlet proposal density.

Step 4: Sample $\tilde{z}_i^{(1)}$ independently from the conditional

$$\tilde{z}_i^{(1)}|\tilde{z}^{(0)}, D^*, \tilde{\Gamma}, y, D \overset{ind}{\sim} \begin{cases} TN_{(\tilde{\alpha}^{(1)}_{y_i}, \tilde{\alpha}^{(1)}_{y_i+1})}(\tilde{\mu}^c_{i1}, \tilde{\sigma}^c_1) & \text{if } D_i = 1 \\ N(\tilde{\mu}^c_{i1}, \tilde{\sigma}^c_1) & \text{if } D_i = 0 \end{cases}, \quad i = 1, 2, \cdots, n.$$

This is a Normal density with mean $\tilde{\mu}^c_{i1}$ and variance $\tilde{\sigma}^c_1$, and is truncated to the interval $(\tilde{\alpha}^{(1)}_{y_i}, \tilde{\alpha}^{(1)}_{y_i+1})$ if individual i is observed to be in the treatment group. When $D_i = 0$, no

[13] Note that sampling the cutpoints in this way enforces the ordering restriction on the cutpoint values.

restrictions arise regarding the latent data $z_i^{(1)}$, and thus the draw is obtained from the untruncated Normal density.

To generate draws from a univariate truncated Normal density, one can use standard inversion methods. That is, to generate $x \sim TN_{(a,b)}(\mu, \sigma^2)$, first draw U uniformly on $(0, 1)$ and then set

$$x = \mu + \sigma\Phi^{-1}\left[\Phi\left(\frac{a-\mu}{\sigma}\right) + U\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]\right].$$

Step 5: By similar arguments as those leading up to step 3, one can show that

$$\tilde{\alpha}^{(0)}|\tilde{z}^{(1)}, D^*, \tilde{\Gamma}_{-\tilde{\alpha}^{(0)}}, y, D \propto \prod_{i:D_i=0} \Phi\left(\frac{\tilde{\alpha}^{(0)}_{y_i+1} - \tilde{\mu}^c_{i0}}{\sqrt{\tilde{\sigma}^c_0}}\right) - \Phi\left(\frac{\tilde{\alpha}^{(0)}_{y_i} - \tilde{\mu}^c_{i0}}{\sqrt{\tilde{\sigma}^c_0}}\right) \tag{24}$$

where

$$\tilde{\mu}^c_{i0} \equiv x_i\tilde{\beta}^{(0)} + [\tilde{\sigma}_{0D}\ \tilde{\sigma}_{10}]\begin{bmatrix} 1 & \tilde{\sigma}_{1D} \\ \tilde{\sigma}_{1D} & \sigma_1 \end{bmatrix}^{-1}\begin{bmatrix} D_i^* - w_i\beta^{(D)} \\ \tilde{z}_i^{(1)} - x_i\tilde{\beta}^{(1)} \end{bmatrix}$$

and

$$\tilde{\sigma}^c_0 \equiv \sigma_0 - [\tilde{\sigma}_{0D}\ \tilde{\sigma}_{10}]\begin{bmatrix} 1 & \tilde{\sigma}_{1D} \\ \tilde{\sigma}_{1D} & \sigma_1 \end{bmatrix}^{-1}\begin{bmatrix} \tilde{\sigma}_{0D} \\ \tilde{\sigma}_{10} \end{bmatrix}.$$

A strategy identical to that described in Step 3 can be used to simulate the cutpoints from this proposal density.

Step 6: Sample $\tilde{z}_i^{(0)}$ independently from the conditional

$$\tilde{z}_i^{(0)}|\tilde{z}^{(1)}, D^*, \tilde{\Gamma}, y, D \overset{ind}{\sim} \begin{cases} TN_{(\tilde{\alpha}^{(0)}_{y_i}, \tilde{\alpha}^{(0)}_{y_i+1})}\,(\tilde{\mu}^c_{i0}, \tilde{\sigma}^c_0) & \text{if } D_i = 0 \\ N(\tilde{\mu}^c_{i0}, \tilde{\sigma}^c_0) & \text{if } D_i = 1 \end{cases}, \quad i = 1, 2, \cdots, n.$$

Step 7: Sample D_i^* independently from the conditional

$$D_i^*|\tilde{z}^0, \tilde{z}^{(1)}, \tilde{\Gamma}, y, D \overset{ind}{\sim} \begin{cases} TN_{(0,\infty)}\,(\tilde{\mu}^c_{iD}, \tilde{\sigma}^c_D) & \text{if } D_i = 1 \\ TN_{(-\infty,0)}\,(\tilde{\mu}^c_{iD}, \tilde{\sigma}^c_D) & \text{if } D_i = 0 \end{cases}, \quad i = 1, 2, \cdots, n.$$

In the above, we have defined

$$\tilde{\mu}^c_{iD} \equiv w_i\beta^{(D)} + [\tilde{\sigma}_{1D}\ \tilde{\sigma}_{0D}]\begin{bmatrix} \sigma_1 & \tilde{\sigma}_{10} \\ \tilde{\sigma}_{10} & \sigma_0 \end{bmatrix}^{-1}\begin{bmatrix} \tilde{z}_i^{(1)} - x_i\tilde{\beta}^{(1)} \\ \tilde{z}_i^{(0)} - x_i\tilde{\beta}^{(0)} \end{bmatrix}$$

and

$$\tilde{\sigma}^c_D \equiv 1 - [\tilde{\sigma}_{1D}\ \tilde{\sigma}_{0D}]\begin{bmatrix} \sigma_1 & \tilde{\sigma}_{10} \\ \tilde{\sigma}_{10} & \sigma_0 \end{bmatrix}^{-1}\begin{bmatrix} \tilde{\sigma}_{1D} \\ \tilde{\sigma}_{0D} \end{bmatrix}.$$

Iterating through steps 1-7 produces a draw from the augmented joint posterior distribution. To recover the structural coefficients of interest, we simply "invert" the mappings described above (13) and below (17).

3.4 Extending the Model: Allowing for Non-Normality

A limitation of the model described thus far in this paper is its reliance on joint Normality. For some applications, such as log wage outcomes [e.g., Heckman and Sedlacek (1985), Heckman (2004)], the Normality assumption may be a reasonable approximation, and if the model passes a selection of diagnostic tests[14] no further refinements would be required. For other models, researchers may worry about heavy tails, asymmetry or possibly bimodality in the disturbance variance. Below we outline simple computational tricks for capturing these features of the data and generalizing the Normality assumption.

The most straight-forward extension of the model is to expand to Student-t errors by simply adding the appropriate mixing variables to the disturbance variance [see, e.g., Carlin and Polson (1991), Geweke (1993), Albert and Chib (1993), Chib and Hamilton (2000) and Li, Poirier and Tobias (2004)]. For example, if we generalize the Normality assumption in (6) to

$$\begin{bmatrix} u_i \\ \epsilon_i^{(1)} \\ \epsilon_i^{(0)} \end{bmatrix} |\lambda_i, x_i, w_i, \Sigma \overset{ind}{\sim} N\left(0, \lambda_i \Sigma\right), \quad \text{where } \Sigma \equiv \begin{bmatrix} 1 & \rho^{(1)} & \rho^{(0)} \\ \rho^{(1)} & 1 & \rho^{(10)} \\ \rho^{(0)} & \rho^{(10)} & 1 \end{bmatrix}, \tag{25}$$

and specify a prior for λ_i of the form[15]

$$\lambda_i \overset{iid}{\sim} IG(\nu/2, 2/\nu),$$

it follows that (marginalized over the prior for λ_i):

$$\begin{bmatrix} u_i \\ \epsilon_i^{(1)} \\ \epsilon_i^{(0)} \end{bmatrix} |x_i, w_i, \Sigma \sim t_\nu(0, \Sigma), \tag{26}$$

a multivariate Student-t density with mean zero, scale matrix Σ and ν degrees of freedom. This device is particularly useful for modeling symmetric error densities whose tails are heavier than those implied by the Normal density. In addition, such an extension to the model comes at little computational cost since, *conditioned on* $\{\lambda_i\}$, sampling the regression parameters and covariance matrix is straight-forward, and each λ_i can be drawn independently from its complete posterior conditional, which is of an inverse Gamma form.

[14] For example, one can calculate posterior predictive p-values [Gelman et al (2004)], QQ plots and other standard diagnostic criteria [e.g., Lancaster (2004), Koop, Poirier and Tobias (2006)] to evaluate the appropriateness of the Normality assumption. For more on the performance of related models under non-Normality, see, for example Goldberger (1983) or Paarsch (1984).

[15] The inverted Gamma (IG) random variable is parameterized as follows: $p(x) \propto x^{-(a+1)} \exp[-1/(bx)]$.

An analogous and potentially more flexible extension of the model is to suppose that the errors were drawn from a mixture of Normal densities. Like (25), we might write

$$\begin{bmatrix} u_i \\ \epsilon_i^{(1)} \\ \epsilon_i^{(0)} \end{bmatrix} |\lambda_i, x_i, w_i, \Sigma \overset{ind}{\sim} \lambda_i N(0, \Sigma_1) + (1-\lambda_i) N(0, \Sigma_2). \tag{27}$$

So, conditioned on λ_i (which is unobserved), each observation is ascribed to one component of the mixture model with covariance matrix equal to either Σ_1 or Σ_2. Since the component assignment is known given λ_i, it is, again, straight-forward to obtain draws from the regression parameters and component-specific covariance matrices. The λ_i are then simulated independently from a two-point distribution.[16]

Finally, one can generalize this mixture model even further by allowing the regression parameters to vary across the mixture components. To do this, we write

$$s_i^* | \lambda_i, \Gamma \overset{ind}{\sim} \lambda_i N(r_i \beta_1, \Sigma_1) + (1-\lambda_i) N(r_i \beta_2, \Sigma_2),$$

where s_i and r_i are as defined in (12), and β_j, Σ_j represent the regression parameter vector and covariance matrix from the j^{th} component of the mixture. Generalization to more than two components is a also straightforward, and the component indicators and component probabilities can be simulated from multinomial and Dirichlet densities, respectively [see, e.g., Li, Poirier and Tobias (2004)].

4 Treatment Effects

In this section we derive expressions for conventional "treatment effects" in our ordered outcome treatment-response model. In particular, we adapt conventional treatment parameters including the Average Treatment Effect (ATE), the effect of Treatment on the Treated (TT) and the Local Average Treatment Effect (LATE) to our ordered response model, and describe how these can be calculated within this framework.

[16] See., e.g., McLachlan and Peel (2000). There is an important issue about local non-identifiability of the mixture model parameters; the parameters are not identified up to a permutation of the mixture components. To aid in identification, priors can be used that impose an ordering restriction on the variance parameters, regression parameters or component probabilities. In some cases, there is little concern for "component switching," but in other cases, this issue may be a significant concern.

4.1 The Average Treatment Effect

We begin with a discussion of the *Average Treatment Effect* (ATE). This parameter typically quantifies the expected outcome gain for a randomly chosen individual. Since our response is ordered, this parameter may not be of direct relevance, as it demands a cardinal representation of an ordinal variable.[17] In light of this issue, we choose to adapt the ATE parameter to describe across-regime changes in probabilities associated with various categories. To fix ideas, then, we consider the impact of the treatment on increasing (or decreasing) the probability that the outcome exceeds the "lowest" category:

$$\begin{aligned} ATE(x;\Gamma) &\equiv \Pr(y^{(1)} \geq 2|x,\Gamma) - \Pr(y^{(0)} \geq 2|x,\Gamma) && (28)\\ &= \sum_{j=2}^{J}\Big[\Pr(y^{(1)} = j|x,\Gamma) - \Pr(y^{(0)} = j|x,\Gamma)\Big] && (29)\\ &= \sum_{j=2}^{J}\Big(\Big[\Phi(\alpha_{j+1}^{(1)} - x\beta^{(1)}) - \Phi(\alpha_{j}^{(1)} - x\beta^{(1)})\Big] - \Big[\Phi(\alpha_{j+1}^{(0)} - x\beta^{(0)}) - \Phi(\alpha_{j}^{(0)} - x\beta^{(0)})\Big]\Big). \end{aligned}$$

The choice of the lowest category is without loss of generality; other probabilities can be obtained in similar ways. We relate this quantity to ATE since it corresponds to a probability increase (or decrease) for a randomly chosen individual.

A point estimate of this treatment impact is readily obtained using our simulated set of parameters drawn from the joint posterior:

$$A\hat{T}E(x) \equiv E_{\Gamma|y,D}\left[ATE(x;\Gamma)\right] \approx \frac{1}{M}\sum_{m=1}^{M} ATE(x;\Gamma_m), \qquad (30)$$

where $\Gamma_m \sim p(\Gamma|y,D)$ and is obtained from the algorithm described in section 3.3.

4.2 The Effect of Treatment on the Treated

The effect of *Treatment on the Treated* (TT) is a conceptually different parameter and describes the outcome gain (or loss) from treatment for those actually selecting into treatment.

[17] In some cases, however, it may be. For example, one could use an ordered model to analyze, say, years of schooling completed, and thus remain true to the integer-valued nature of the education data. In this case, the ordered variable has a natural cardinal interpretation, and thus the conventional ATE parameter would be of interest.

Again, we examine the treatment effect on the probability that the outcome variable does not fall into the lowest category:

$$\begin{aligned} TT(x,w,D(w)=1;\Gamma) &\equiv \Pr(y^{(1)} \geq 2|x,w,D(w)=1,\Gamma) - \Pr(y^{(0)} \geq 2|x,w,D(w)=1,\Gamma) \quad (31) \\ &= \sum_{j=2}^{J} \Big[\Pr(y^{(1)} = j|x,w,D(w)=1,\Gamma) - \Pr(y^{(0)} = j|x,w,D(w)=1,\Gamma)\Big]. \end{aligned}$$

To economize on notation, let us define

$$P_{1,j}^{TT}(\Gamma) \equiv \Pr(y^{(1)} = j|x,w,D(w)=1,\Gamma) \quad \text{and} \quad P_{0,j}^{TT}(\Gamma) \equiv \Pr(y^{(0)} = j|x,w,D(w)=1,\Gamma), \tag{32}$$

keeping the conditioning on x, w and $D(w)=1$ implicit. Given these definitions, it follows that $TT(x,w,D(w)=1;\Gamma) = \sum_{j=2}^{J}[P_{1,j}^{TT}(\Gamma) - P_{0,j}^{TT}(\Gamma)]$.

Recalling our description of the likelihood in section 2.1, we can write the probabilities in (32) in more computationally convenient forms. For example,

$$\begin{aligned} P_{1,j}^{TT}(\Gamma) &\equiv \Pr(\alpha_j^{(1)} < z^{(1)} \leq \alpha_{j+1}^{(1)}|u > -w\beta^{(D)}) \\ &= \Pr(\alpha_j^{(1)} - x\beta^{(1)} < \epsilon^{(1)} \leq \alpha_{j+1}^{(1)} - x\beta^{(1)}|u > -w\beta^{(D)}) \\ &= \int_{\alpha_j^{(1)}-x\beta^{(1)}}^{\alpha_{j+1}^{(1)}-x\beta^{(1)}} p(\epsilon^{(1)}|u > -w\beta^{(D)})d\epsilon^{(1)} \\ &= \int_{\alpha_j^{(1)}-x\beta^{(1)}}^{\alpha_{j+1}^{(1)}-x\beta^{(1)}} \int_{-w\beta^{(D)}}^{\infty} \frac{p(\epsilon^{(1)},u)}{\Pr(u > -w\beta^{(D)})} du d\epsilon^{(1)} \\ &= \int_{-w\beta^{(D)}}^{\infty} \int_{\alpha_j^{(1)}-x\beta^{(1)}}^{\alpha_{j+1}^{(1)}-x\beta^{(1)}} p(\epsilon^{(1)}|u)d\epsilon^{(1)} \frac{p(u)}{\Pr(u > -w\beta^{(D)})} du \\ &= \int_{-w\beta^{(D)}}^{\infty} \left[\Phi\left(\frac{\alpha_{j+1}^{(1)} - x\beta^{(1)} - \rho^{(1)}u}{\sqrt{1-\rho^{(1)2}}}\right) - \Phi\left(\frac{\alpha_j^{(1)} - x\beta^{(1)} - \rho^{(1)}u}{\sqrt{1-\rho^{(1)2}}}\right)\right] \frac{p(u)}{\Pr(u > -w\beta^{(D)})} du. \end{aligned}$$

The integral above is simply

$$E_u\left[\Phi\left(\frac{\alpha_{j+1}^{(1)} - x\beta^{(1)} - \rho^{(1)}u}{\sqrt{1-\rho^{(1)2}}}\right) - \Phi\left(\frac{\alpha_j^{(1)} - x\beta^{(1)} - \rho^{(1)}u}{\sqrt{1-\rho^{(1)2}}}\right)\right],$$

where $u \sim TN_{(-w\beta^{(D)},\infty)}(0,1)$. Thus, the strong law of large numbers guarantees that

$$\hat{P}_{1,j}^{TT}(\Gamma) \equiv (1/L)\sum_{l=1}^{L} \Phi\left(\frac{\alpha_{j+1}^{(1)} - x\beta^{(1)} - \rho^{(1)}u^{(l)}}{\sqrt{1-\rho^{(1)2}}}\right) - \Phi\left(\frac{\alpha_j^{(1)} - x\beta^{(1)} - \rho^{(1)}u^{(l)}}{\sqrt{1-\rho^{(1)2}}}\right) \xrightarrow{p} P_{1,j}^{TT}(\Gamma),$$

where $\{u^{(l)}\}_{l=1}^{L}$ denotes an iid sample from the standard Normal distribution truncated to $(-w\beta^{(D)}, \infty)$.[18]

Following similar arguments, one can show

$$\hat{P}_{0,j}^{TT}(\Gamma) \equiv (1/L)\sum_{l=1}^{L} \Phi\left(\frac{\alpha_{j+1}^{(0)} - x\beta^{(0)} - \rho^{(0)}u^{(l)}}{\sqrt{1-\rho^{(0)^2}}}\right) - \Phi\left(\frac{\alpha_{j}^{(0)} - x\beta^{(0)} - \rho^{(0)}u^{(l)}}{\sqrt{1-\rho^{(0)^2}}}\right) \xrightarrow{p} P_{0,j}^{TT}(\Gamma).$$

Putting these results together, and, of course, exploiting the availability of draws from our joint posterior, we can calculate the following point estimate of TT:

$$\begin{aligned} \hat{TT}(x, w, D(w) = 1) &= E_{\Gamma|y,D}[TT(x, w, D(w) = 1; \Gamma)] \\ &\approx \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{j=2}^{J}\left(\hat{P}_{1,j}^{TT}(\Gamma_m) - \hat{P}_{0,j}^{TT}(\Gamma_m)\right)\right], \end{aligned} \tag{33}$$

with $\Gamma_m \sim p(\Gamma|y, D)$.

4.3 The Local Average Treatment Effect

The *Local Average Treatment Effect* can be interpreted as measuring the outcome gain (or loss) from treatment for a group of "compliers." This corresponds to the effect of treatment on a subgroup of the population who would choose to receive treatment at a particular value of the instrument, say w, but would not choose treatment at some $\tilde{w}$.[19] Consistent with our discussions in the previous subsections, our parameter of interest is the increased (or decreased) likelihood that the outcome variable exceeds the lowest category:

$$\begin{aligned} LATE(x, w, \tilde{w}, D(w) = 1, D(\tilde{w}) = 0; \Gamma) &= \Pr(y^{(1)} \geq 2|x, w, \tilde{w}, D(w) = 1, D(\tilde{w}) = 0, \Gamma) \\ &- \Pr(y^{(0)} \geq 2|x, w, \tilde{w}, D(w) = 1, D(\tilde{w}) = 0, \Gamma) \\ &= \sum_{j=2}^{J}\left(P_{1,j}^{LATE}(\Gamma) - P_{0,j}^{LATE}(\Gamma)\right), \end{aligned}$$

where

$$P_{k,j}^{LATE}(\Gamma) \equiv \Pr(y^{(k)} = j|x, w, \tilde{w}, D(w) = 1, D(\tilde{w}) = 0, \Gamma), \quad k = 0, 1.$$

[18] In practice, these integrals can be approximated quite accurately (and quickly) using relatively few draws from the truncated Normal distribution. A routine for drawing from such a distribution was provided in Step 4 of section 3.3.

[19] Heckman, Tobias and Vytlacil (2003) provide a similar definition of LATE in a parametric latent variable selection model.

To calculate LATE, we follow a similar strategy to that outlined for calculating the TT effect. It follows that

$$\begin{aligned}\hat{P}^{LATE}_{1,j}(\Gamma) &\equiv (1/L)\sum_{l=1}^{L}\Phi\left(\frac{\alpha^{(1)}_{j+1} - x\beta^{(1)} - \rho^{(1)}u^{(l)}}{\sqrt{1-\rho^{(1)^2}}}\right) - \Phi\left(\frac{\alpha^{(1)}_{j} - x\beta^{(1)} - \rho^{(1)}u^{(l)}}{\sqrt{1-\rho^{(1)^2}}}\right)\\ &\xrightarrow{p} P^{LATE}_{1,j}(\Gamma),\end{aligned}$$

$$\begin{aligned}\hat{P}^{LATE}_{0,j}(\Gamma) &\equiv (1/L)\sum_{l=1}^{L}\Phi\left(\frac{\alpha^{(0)}_{j+1} - x\beta^{(0)} - \rho^{(0)}u^{(l)}}{\sqrt{1-\rho^{(0)^2}}}\right) - \Phi\left(\frac{\alpha^{(0)}_{j} - x\beta^{(0)} - \rho^{(0)}u^{(l)}}{\sqrt{1-\rho^{(0)^2}}}\right)\\ &\xrightarrow{p} P^{LATE}_{0,j}(\Gamma),\end{aligned}$$

where $u^{(l)} \overset{iid}{\sim} TN_{[-w\beta^{(D)},-\tilde{w}\beta^{(D)}]}(0,1)$. We can then proceed to obtain a point estimate of LATE:

$$\begin{aligned}L\hat{AT}E[x,w,\tilde{w},D(w)=1,D(\tilde{w}=0)] &= E_{\Gamma|y,D}[LATE(x,w,\tilde{w},D(w)=1,D(\tilde{w})=0;\Gamma)]\\ &\approx \frac{1}{M}\sum_{m=1}^{M}\left[\sum_{j=2}^{J}\left(\hat{P}^{LATE}_{1,j}(\Gamma_m) - \hat{P}^{LATE}_{0,j}(\Gamma_m)\right)\right],\end{aligned}$$

with $\Gamma_m \sim p(\Gamma|y,D)$.

4.4 Beyond Mean Treatment Parameters: Learning about $\rho^{(10)}$

The treatment parameters discussed in the previous subsections are typical of the mean treatment effects considered in the literature. To see this, note that an equivalent expression of the parameter of interest $\Pr(y^{(1)} \geq 2|x) - \Pr(y^{(0)} \geq 2|x)$ is $E[I(y^{(1)} \geq 2) - I(y^{(0)} \geq 2)|x]$, where $I(\cdot)$ denotes the indicator function. Part of the appeal of these mean treatment parameters is that they enable researchers to quantify a feature of the treatment impact - the average gains or losses under various conditioning scenarios - even though $y^{(0)}$ and $y^{(1)}$ are not jointly observed. Unlike the mean treatment effects described in the previous subsections, however, other quantities of significant policy relevance, such as the probability of a positive treatment effect

$$\Pr(y^{(1)} - y^{(0)} > 0|x)$$

will depend on the correlation parameter $\rho^{(10)}$. This correlation parameter does not enter the likelihood for the observed data (see, e.g., section 2.1) and thus *is not identifiable.* This fact has, perhaps, limited the scope of most research to the estimation of mean treatment impacts.[20]

For the Bayesian, this non-identifiability issue raises the question of what can and should be done about our treatment of the correlation parameter $\rho^{(10)}$.[21] One approach, which was used by Chib and Hamilton (2000), is to simply set $\rho^{(10)} = 0$, fit the model subject to this restriction and then impose that the *restricted* covariance matrix (subject to $\rho^{(10)} = 0$) is positive definite. While in most cases this will be an innocuous restriction, in some cases, this approach may have unanticipated consequences. For example, if we set $\rho_{10} = 0$ in (6), it follows that Σ is positive definite if and only if

$$[\rho^{(1)}]^2 + [\rho^{(0)}]^2 \leq 1.$$

This restriction thus forces the *identified* correlation parameters $\rho^{(1)}$ and $\rho^{(0)}$ to lie within the unit circle rather than the unit square. To illustrate what this restriction means, suppose that we performed a generated data experiment and set $\rho^{(1)} = \rho^{(0)} = .8$. If we proceeded to fit this model subject to $\rho^{(10)} = 0$, and enforced that the restricted covariance matrix was positive definite, then our posterior mode must be *inconsistent* - the joint posterior $\rho^{(1)}$, $\rho^{(0)}|y, D$ could never place *any mass* over the actual values used to generate the data regardless of the size of the generated data set. This problem manifests itself for rather extreme cases of correlation among the unobservables; if the correlations are more moderate, then this is not likely to be a significant issue.[22]

An alternate approach, which we have advocated in previous work [e.g., Koop and Poirier (1997), Poirier and Tobias (2003), Li, Poirier and Tobias (2004)] is to simply work with the "full" covariance matrix, as described in section 3.3, without restricting $\rho^{(10)}$ *a priori.* As shown in Poirier and Tobias (2003), this does not induce an inconsistency regarding the identified model parameters, and moreover, *one can potentially learn about the non-identified correlation parameter.* Intuitively, information arising through the likelihood function will enable us to pin down all of the correlation parameters in (6) that are identifiable, leaving

[20] Related work has sought to expand the focus beyond mean effects and identify outcome gain distributions. See, for example, Heckman and Honoré (1990), Heckman, Smith and Clements (1997), Heckman and Smith (1998) and Carneiro, Hansen and Heckman (2003).

[21] See Poirier (1998) for more on learning about non-identifiable parameters through prior information. Poirier and Tobias (2000) contain related material describing the implications of prior restrictions on $\rho^{(10)}$.

[22] This is particularly true, as Chib and Hamilton (2000) point out, in, say, panel models where most of the variation is captured through fixed or random effects, and one would suspect that any remaining correlation among the unobservables was minimal.

only $\rho^{(10)}$ unknown. An additional source of information then arises from the fact that Σ must be positive definite. In particular, the p.d. restriction imposes that $\rho^{(10)}$ must have the following conditional support:

$$\rho^{(1)}\rho^{(0)} - \sqrt{(1-[\rho^{(1)}]^2)(1-[\rho^{(0)}]^2)} \leq \rho^{(10)} \leq \rho^{(1)}\rho^{(0)} + \sqrt{(1-[\rho^{(1)}]^2)(1-[\rho^{(0)}]^2)}. \tag{34}$$

This equation provides identifiable bounds on $\rho^{(10)}$ as a function of the identified correlation parameters $\rho^{(1)}$ and $\rho^{(0)}$. *Somewhat surprisingly, these bounds also suggest that selection bias may, in a particular sense, be a good thing.* When $\rho^{(1)}$ and $\rho^{(0)}$ are large, the bounds given in (34) become increasingly informative. Intuitively, the presence of selection bias provides a vehicle for learning about $\rho^{(10)}$ - if the errors in the outcome equations are correlated sufficiently with the error in the treatment equation, then to some extent, they must also be correlated with one another.

Equation (34) shows that beliefs regarding $\rho^{(10)}$ will generally be revised from the data - as we learn about $\rho^{(1)}$ and $\rho^{(0)}$, this information spills over and restricts the conditional support of $\rho^{(10)}$. This is, unfortunately, as far as the data will take us - the shape of the marginal posterior density of $\rho^{(10)}$ within the bounds in (34) is not updated from the data. Poirier and Tobias (2003) show that in sufficiently large samples where $\rho^{(1)}$ and $\rho^{(0)}$ are estimated precisely and are approximately equal to, say, $\rho^{(1,*)}$ and $\rho^{(0,*)}$:

$$p(\rho^{(10)}|y, D) \approx p(\rho^{(10)}|\rho^{(1)} = \rho^{(1,*)}, \rho^{(0)} = \rho^{(0,*)}). \tag{35}$$

That is, *the marginal posterior for the non-identified correlation parameter is approximately equal to the conditional prior for that correlation parameter evaluated at the given values $\rho^{(1,*)}$ and $\rho^{(0,*)}$* . The support bounds in (34) are updated from the data, but within the bounds, the shape of the posterior is completely determined by the shape of the conditional prior. For the Bayesian, this is a natural result; in the absence of information arising from the data, one resorts to the use of prior information.[23]

The results of these studies suggest that there is, in one sense, a limited opportunity for expanding the focus of research beyond mean effects. One could at least bound $\rho^{(10)}$ and then use these bounds to bound other parameters of interest. If one is comfortable with

[23] Heckman, Smith and Clements (1997), for example, informally discuss plausible prior beliefs for $\rho^{(10)}$. They write (page 510) "In considering outcomes like employment and earnings, many plausible models of program participation suggest that outcomes in the treatment state are "positively related" to outcomes in the non-treatment state...There is a widely-held belief that good persons are good at whatever they do."

insinuating prior information, however, one could obtain point estimates of any parameter of interest under a particular prior. "Default" priors yielding marginal posteriors that are uniform over the conditional support bounds may appeal to many researchers when carrying out these calculations. Use of such priors, however, typically makes the problem more challenging from a computational point of view as they often break the inherent conjugacy of the model.

5 A Generated Data Experiment

In this section we conduct a generated data experiment to demonstrate the performance of our posterior simulator and address a potential concern regarding choice of prior. A sample of 5,000 observations is generated from the following ordered potential outcome model:

$$\begin{aligned} D_i^* &= \beta_0^{(D)} + w_i \beta_1^{(D)} + u_i \\ z_i^{(1)} &= \beta^{(1)} + \epsilon_i^{(1)} \\ z_i^{(0)} &= \beta^{(0)} + \epsilon_i^{(0)}, \end{aligned}$$

where w_i is drawn independently from a $N(0,1)$ distribution and the error terms $[u_i\ \epsilon_i^{(1)}\ \epsilon_i^{(0)}]'$ are drawn jointly from the trivariate Normal distribution:

$$\left[\begin{array}{c} u_i \\ \epsilon_i^{(1)} \\ \epsilon_i^{(0)} \end{array} \right] | w_i \overset{iid}{\sim} N \left[\left(\begin{array}{c} 0 \\ 0 \\ 0 \end{array} \right), \left(\begin{array}{ccc} 1 & 0.9 & 0.7 \\ 0.9 & 1 & 0.6 \\ 0.7 & 0.6 & 1 \end{array} \right) \right].$$

We consider this specific design with a high degree of unobservable correlation to reveal how our algorithm performs when selection bias is a significant problem.[24] The non-identified correlation $\rho^{(10)}$ is set to .6, and thus from (34) the covariance matrix is positive definite. Finally, the regression parameters $\beta_0^{(D)}$, $\beta_1^{(D)}$, $\beta^{(1)}$ and $\beta^{(0)}$ and cutpoint values $\alpha_j^{(k)}$, $j = 3, 4, 5$, $k = 0, 1$ are enumerated in the first column of Table 1, and the observables D_i, $y_i^{(1)}$ and $y_i^{(0)}$ are generated as follows:

$$\begin{aligned} D_i &= I(D_i^* > 0), \\ y_i^{(1)} &= j \quad \text{if} \quad \alpha_j^{(1)} < z_i^{(1)} \leq \alpha_{j+1}^{(1)}, \quad j = 1, 2, 3, 4, 5, \\ y_i^{(0)} &= j \quad \text{if} \quad \alpha_j^{(0)} < z_i^{(0)} \leq \alpha_{j+1}^{(0)}, \quad j = 1, 2, 3, 4, 5. \end{aligned}$$

[24] We do not address the "weak instruments" problem here, but to fix ideas, we consider the case where the instrument plays a significant role in the treatment decision.

With this experimental design, the number of treated versus untreated observations is well-balanced, as 51% of the sample points are assigned to the treatment group. Of those sample points that are assigned to the treatment group, 5%, 5%, 8%, 10% and 71% are associated with ordered outcomes of $y^{(1)} = 1, 2, 3, 4$ and 5, respectively. Likewise, for those observations that do not receive treatment, 46%, 14%, 13%, 10% and 16% of them fall into the categories of $y^{(0)} = 1, 2, 3, 4$ and 5, respectively. We consider this design to be reasonably typical of actual empirical situations, where the outcome variables are not uniformly distributed over the set of possible choices.

We fit our model using the posterior simulator described in section 3.3, run the algorithm for 3,000 iterations, and discard the first 600 draws as the burn-in period. To illustrate the performance of the algorithm, we plot in Figure 1 the *lagged autocorrelations* up to order 100 for several selected parameters: $\beta_0^{(D)}, \beta^{(1)}, \alpha_3^{(0)}$ and $\rho^{(1)}$. The lagged autocorrelation plots are a useful way to assess the mixing of the parameter chains - if the lagged autocorrelations remain close to unity, for example, then the posterior simulator only makes small local movements from iteration to iteration, resulting in inaccurate posterior estimates. As shown in Figure 1, the lagged autocorrelations drop away reasonably quickly for all the selected parameters, suggesting that posterior quantities can be approximated reasonably accurately with only a moderate number of posterior simulations.

$\star$ Figure 1 about here $\star$

As discussed in section 3.2, one potential concern about working with the reparameterized model is that we need to impose priors directly on the transformed parameters instead of the structural parameters. This is an important issue because priors that look suitable for the transformed parameters may turn out to imply rather unreasonable (and possibly quite informative) priors for the structural parameters. For this generated data experiment, we employ the priors described in Section 3.1. We can calculate the implied priors for the structural parameters by first sampling from the priors for the transformed parameters, inverting to obtain the values of the structural parameters, and then smoothing the collection of structural parameter values to obtain their approximate marginal prior densities. To demonstrate this process, we plot in Figure 2 the marginal priors and posteriors for the selected parameters $\beta_0^{(D)}, \beta^{(1)}, \alpha_3^{(0)}$ and $\rho^{(1)}$. As can be seen clearly from the graphs, the prior densities for all the parameters are almost completely “flat” over the regions where the posterior densities have substantial mass; it is almost impossible for us to visually distinguish between the prior densities and horizontal axes when they are plotted together against the posterior densities. This evidence clearly suggests that the data has provided sufficient information for us to substantially revise our prior beliefs, and that the implied priors for the structural coefficients are sufficiently vague to warrant working with the reparameterized model specification.

⋆ Figure 2 about here ⋆

In Table 1, we report posterior estimates of all the parameters along with their true generated data values. As is evident from the table, all the parameters have been estimated with reasonable accuracy and posterior estimates are quite close to their actual values. As for the non-identified correlation parameter $\rho^{(10)}$, its marginal posterior places considerable mass over the actual value that was used to generate the data, .6. We interpret this finding with substantial caution, however, as our point estimate (e.g., posterior mode) of this parameter is *not consistent*; there is absolutely no way that we can recover the "true" value of this parameter even in the largest data sets. What is true, however, is that as we learn about the identified $\rho^{(1)}$ and $\rho^{(0)}$ correlation parameters, (34) restricts the conditional support of $\rho^{(10)}$. For this experimental design, the conditional support bounds for $\rho^{(10)}$ are 0.32 and 0.94, suggesting that there is significant potential for learning about $\rho^{(10)}$. This learning is manifested in the marginal posterior for $\rho^{(10)}$, as the posterior simulations automatically incorporate this support restriction. The fact that most of our posterior mass is centered around the approximate midpoint of this region (.6) is not informative, and is simply a consequence of the shape of our particular prior density. All the data can do is to reveal the support bounds; beyond this, the prior takes over and affects the shape of the posterior within the support bounds.

⋆ Table 1 about here ⋆

To examine if our results are sensitive to alternate prior specifications for the non-identified correlation parameter, we re-estimated our model by specifying a different value of $\underline{R}$ used in the inverted Wishart prior for the covariance matrix $\tilde{\Sigma}$. Specifically, to reflect the potential prior preference for positive $\rho^{(10)}$, we changed the $(2,3)$ and $(3,2)$ elements of $\underline{R}$ to 0.5 (which were formerly zero). Interestingly, and as expected from a theoretical point of view, with this many observations, the identified correlation parameters are not affected by this change in prior, and thus the bounds described in (34) continue to be pinned down rather precisely. The shape of the prior within these bounds does change, however, as the shift toward positive values increases the marginal posterior mean to 0.742 instead of 0.635, as described in the previous table.

In the second part of Table 1 we also list the true values and posterior estimates of the mean treatment effects ATE, TT and LATE. As discussed in Section 4, these quantities summarize the impact of the treatment in increasing (or decreasing) the probability that the outcome exceeds the lowest category. When calculating TT and LATE (which are functions of covariates in the treatment decision), we set $w = 0$ for TT, and for LATE, set $w = 0$ and $\tilde{w} = -1$. In the bottom portion of Table 1 we also illustrate how parameters beyond mean effects, such as the probabilities of positive, negative or zero treatment impacts for various subpopulations, can be calculated. The fact that the conventional mean treatment parameters are pinned down quite accurately in Table 1 is not surprising - these effects are purely functions of identified model parameters which are precisely estimated by our simulation algorithm. For the treatment parameters involving the probabilities of positive, negative or zero treatment impacts, we are able to derive reasonably tight bounds around $\rho^{(10)}$, and therefore are able to obtain reasonably accurate estimates of the non-identified quantities of interest. We continue to stress, however, that it is not possible to consistently estimate these quantities, though one could potentially bound them, or as done in this section, one can use prior information to fill the gaps created by the absence of data information.

6 Conclusion

In this paper we introduced a new Bayesian estimation algorithm for fitting a binary treatment, ordered outcome selection model in a potential outcomes framework. Our particular algorithm made use of a reparameterization, building of the suggestion of Nandram and Chen (1996), to accelerate the convergence of our posterior simulator and mitigate problems of slow-mixing. Several computational strategies which allowed for non-Normality were also discussed and conventional “treatment effects” such as the Average Treatment Effect

(ATE), the effect of Treatment on the Treated (TT) and the Local Average Treatment Effect (LATE) were derived for this specific model. We also reviewed how a Bayesian might attempt to expand the focus of her research beyond mean treatment impacts by exploiting a limited degree of learning that takes place about the non-identified cross regime correlation parameter.

References

[1] Angrist, J.D. and A.B. Krueger, (2001), "Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments," *Journal of Economic Perspectives*, 15(4), 69-85.

[2] Albert, J. and S. Chib, (1993), "Bayesian analysis of binary and polychotomous response data," *Journal of the American Statistical Association*, 88, 669-679.

[3] Carlin, B.P. and T.A. Louis, (2000), *Bayes and Empirical Bayes Methods for Data Analysis*, 2nd ed., (Chapman & Hall / CRC).

[4] Carlin, B. and Polson, N., (1991), "Inference for nonconjugate Bayesian models using the Gibbs sampler," *Canadian Journal of Statistics,* 19, 399-405.

[5] Carneiro, P., K. Hansen and J. Heckman, (2003), "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44(2), 361-422.

[6] Casella, G. and E.I. George, (1992), "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167-174.

[7] Chen, M.-H., Q.-M. Shao and J.G. Ibrahim, (2000), *Monte Carlo Methods in Bayesian Computation*, (Springer-Verlag).

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

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

[10] Chib, S. and B.H. Hamilton, (2000), "Bayesian analysis of cross-section and clustered data treatment models," *Journal of Econometrics*, 97(1), 25-50.

[11] Chib, S. and B.H. Hamilton, (2002), "Semiparametric Bayes Analysis of Longitudinal Data Treatment Models," *Journal of Econometrics*, 110, 67-89.

[12] Cowles, M., (1996), "Accelerating Monte Carlo Markov Chain Convergence for Cumulative-Link Generalized Linear Models," *Statistics and Computing*, 6, 101-111.

[13] Gelman, A., J.B. Carlin, H.S. Stern and D.B. Rubin, (2004), *Bayesian Data Analysis*, 2nd ed., (Chapman & Hall / CRC).

[14] Geweke, J., (1993), "Bayesian Treatment of the Independent Student-t Linear Model," *Journal of Applied Econometrics*, S19-S40.

[15] Geweke, J., (1999), "Using Simulation Methods for Bayesian Econometric Models: Inference, Development and Communication (with discussion and reply)," Econometric Reviews, 18, 1-127.

[16] Geweke, J. and M. Keane, (2001), "Computationally Intensive Methods for Integration in Econometrics", in J.J. Heckman and E. Leamer, eds., *Handbook of Econometrics: Volume 5*, (North-Holland: Elsevier Science), 3463-3568.

[17] Gilks, W.R., S. Richardson and D.J. Spiegelhalter, eds., (1998), *Markov Chain Monte Carlo in Practice*, Boca Raton, Fla.: Chapman & Hall/CRC.

[18] Goldberger, A.S., (1983), "Abnormal Selection Bias," in *Studies in Econometrics, Time Series and Multivariate Statistics*, S. Karlin et al, eds. New York: Academic Press.

[19] Gould, E.D., (2002), "Rising wage inequality, comparative advantage, and the growing importance of general skills in the United States," *Journal of Labor Economics*, 20(1), 105-147.

[20] Gould, E.D., (2005), "Inequality and Ability," *Labour Economics*, 12, 169-189.

[21] Heckman, J., (2004), "Micro Data, Heterogeneity and the Evaluation of Public Policy, Part 1," *The American Economist*, 48(2), 3-25.

[22] Heckman, J. and B. Honoré, (1990), "The Empirical Content of the Roy Model," *Econometrica*, 50, 1121-1149.

[23] Heckman, J. and G.L. Sedlacek, (1985), "Heterogeneity, Aggregation and Market Wage Functions: An Empirical Model of Self-Selection in the Labor Market," *Journal of Political Economy*, 93(6), 1077-1125.

[24] Heckman, J. and J. Smith, (1998), "Evaluating the Welfare State," *NBER Working Paper*, 6542, National Bureau of Economic Research, Inc.

[25] Heckman, J. and J. Smith with N. Clements, (1997), "Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts," *Review of Economic Studies*, 64, 487-535.

[26] Heckman, J., J.L. Tobias and E. Vytlacil, (2003), "Simple Estimators for Treatment Parameters in a Latent-Variable Framework," *Review of Economics and Statistics*, 85(3), 748-755.

[27] Heckman, J. and E. Vytlacil, (1999), "Local Instrumental Variables and Latent Variable Models for Identifying and Bounding Treatment Effects," *Proceedings of the National Academy of Sciences*, 96, 4730-4734.

[28] Heckman, J. and E. Vytlacil, (2000), "The Relationship Between Treatment Parameters within a Latent Variable Framework," *Economics Letters*, 33-39.

[29] Imbens, G. and J. Angrist, (1994), "Identification and Estimation of Local Average Treatment Effects," *Econometrica*, 62, 467-475.

[30] Koop, G., (2003), *Bayesian Econometrics*, (Wiley).

[31] Koop, G. and D.J. Poirier, (1997), "Learning About the Across-Regime Correlation in Switching Regression Models," *Journal of Econometrics*, 78, 217-227.

[32] Koop, G., D.J. Poirier, and J.L. Tobias, (2006), *Bayesian Econometrics.* Volume 7 of Econometric Exercises series, Cambridge University Press, forthcoming.

[33] Koop, G. and J.L. Tobias, (2004), "Semiparametric Bayesian Inference in Smooth Coefficient Models," *Journal of Econometrics*, forthcoming.

[34] Lancaster, T., (2004), *An Introduction to Modern Bayesian Econometrics*, (Blackwell).

[35] Li, K., (1998), "Bayesian inference in a simultaneous equation model with limited dependent variables," *Journal of Econometrics*, 85(2), 387-400.

[36] Li, M. and J.L. Tobias, (2005), "Bayesian analysis of structural effects in an ordered equation system," Working Paper, Iowa State University, Department of Economics.

[37] Li, M., D.J. Poirier and J.L. Tobias, (2004), "Do Dropouts Suffer from Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models," *Journal of Applied Econometrics*, 19, 203-225.

[38] Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*, Cambridge, Cambridge University Press.

[39] Manski, C., (1990), "Nonparametric Bounds on Treatment Effects," *American Economic Review Papers and Proceedings*, 80, 319-323.

[40] Manski, C., (1994), "The Selection Problem," in *Advances in Econometrics: Sixth World Congress*, C. Sims(editor), Cambridge, UK: Cambridge University Press, 143-170.

[41] McLachlan, G. and D. Peel, (2000), *Finite Mixture Models*, New York: Wiley.

[42] Munkin, M.K. and P.K. Trivedi, (2003), "Bayesian analysis of a self-selection model with multiple outcomes using simulation-based estimation: An application to the demand for healthcare," *Journal of Econometrics*, 114(2), 197-220.

[43] Nandram, B. and M.-H. Chen, (1996), "Reparameterizing the Generalized linear model to accelerate Gibbs sampler convergence," *Journal of Statistical Computation and Simulation*, 54, 129-144.

[44] Nobile, A., (2000), "Comment: Bayesian Multinomial Probit Models with a Normalization Constraint," *Journal of Econometrics*, 99(2), 335-345.

[45] Paarsch, H. J., (1984), "A Monte Carlo Comparison of Estimators for Censored Regression Models," *Journal of Econometrics*, 24, 197-213.

[46] Poirier, D.J., (1995), *Intermediate Statistics and Econometrics*, (The MIT Press, Cambridge).

[47] Poirier, D.J., (1998), "Revising Beliefs in Non-Identified Models," *Econometric Theory*, 14, 483-509.

[48] Poirier, D.J. and J. L. Tobias, (2000), "Across-Regime Covariance Restrictions in Treatment Response Models," working paper, University of California-Irvine, department of economics.

[49] Poirier, D.J. and J.L. Tobias, (2003), "On the Predictive Distributions of Outcome Gains in the Presence of an Unidentified Parameter," *Journal of Business and Economic Statistics*, 21(2), 258-268.

[50] Poirier, D.J. and J.L. Tobias, (2006), “Bayesian Econometrics,” in *Palgrave Handbook of Econometrics, Volume 1: Theoretical Econometrics* (K. Patterson and T.C. Mills, eds.), Palgrave.

[51] Roy, A. D., (1951), “Some Thoughts on the Distribution of Earnings,” *Oxford Economic Papers*, 3, 135-146.

[52] Tanner, M.A. and W.H. Wong, (1987), “The Calculation of Posterior Distributions by Data Augmentation,” *Journal of the American Statistical Association*, 82, 528-540.

[53] Tierney, L., (1994), “Markov Chains for Exploring Posterior Distributions (with discussion),” *Annals of Statistics*, 22, 1701-1762.

[54] Vijverberg, W.P.M., (1993), “Measuring the Unidentified Parameter of the Extended Roy Model of Selectivity,” *Journal of Econometrics* 57, 69-89.

Figure 1: Lagged autocorrelations of simulated posterior draws for $\beta_0^{(D)}$, $\beta^{(1)}$, $\alpha_3^{(0)}$ and $\rho^{(1)}$

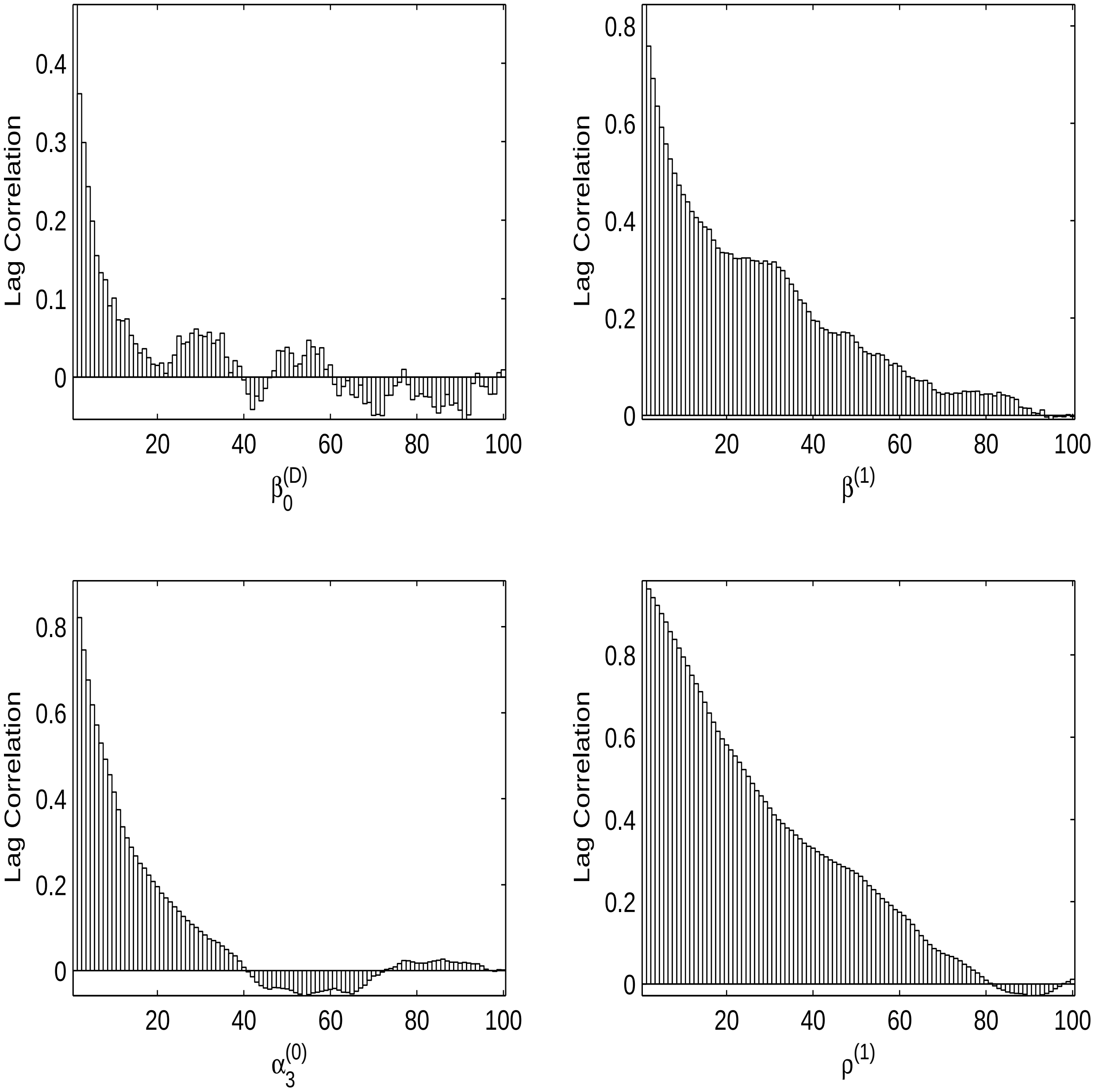

Figure 2: Marginal prior (dashed lines) and posterior (solid lines) density functions for $\beta_0^{(D)}, \beta^{(1)}, \alpha_3^{(0)}$ and $\rho^{(1)}$

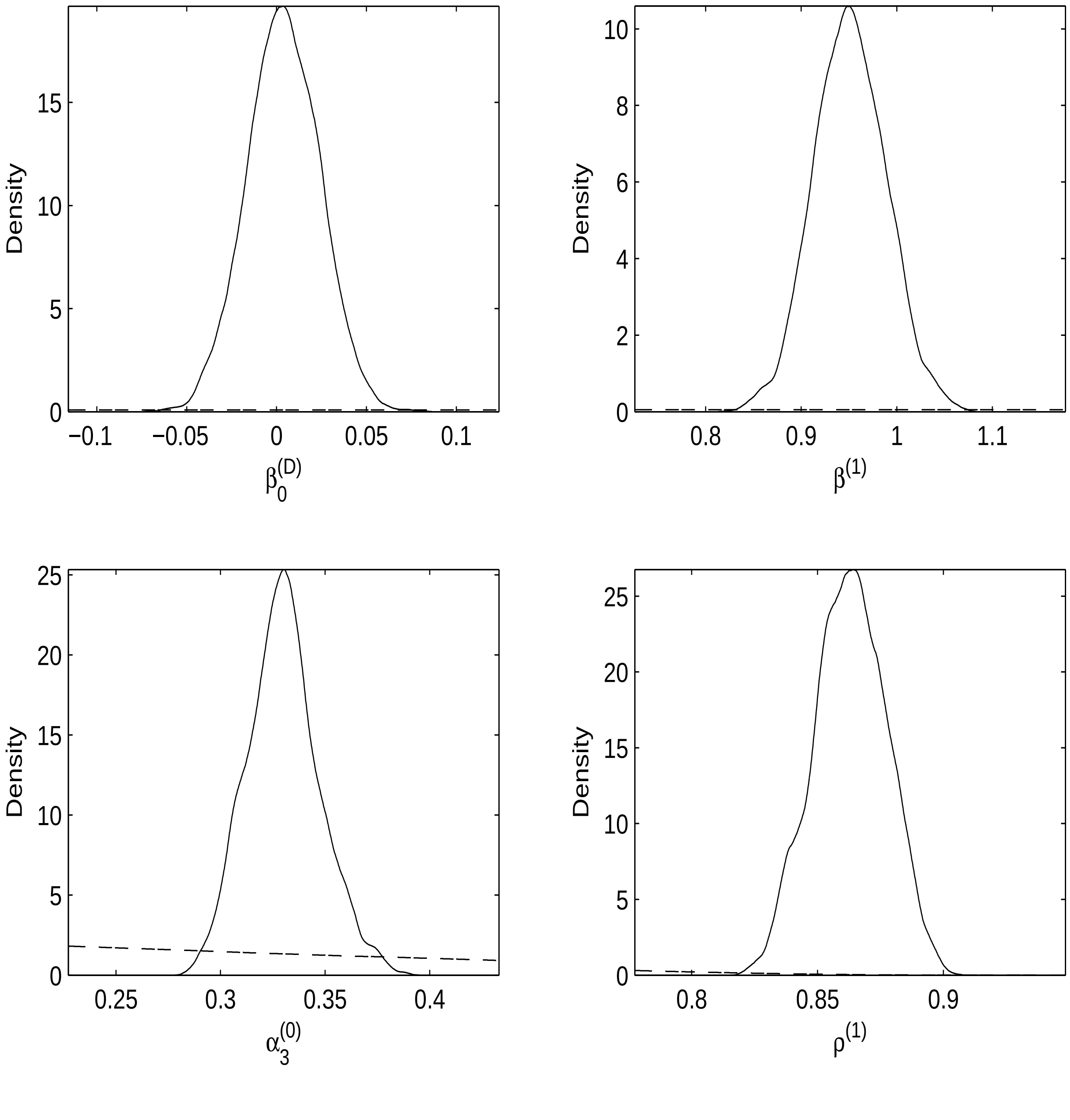

Table 1: True values and posterior estimates of the parameters

	True	Posterior Estimates		
	Values	$\mathrm{E}(\cdot\|D)$	$\mathrm{Std}(\cdot\|D)$	$\mathrm{P}(\cdot > 0\|D)$
Regression Parameters				
$\beta_0^{(D)}$	0	0.0039	0.02	0.576
$\beta_1^{(D)}$	1	1	0.0252	1
$\beta^{(1)}$	0.913	0.951	0.0375	1
$\beta^{(0)}$	0.477	0.468	0.0284	1
Cutpoint Values				
$\alpha_3^{(1)}$	0.304	0.315	0.0226	1
$\alpha_4^{(1)}$	0.609	0.626	0.0267	1
$\alpha_5^{(1)}$	0.913	0.924	0.0298	1
$\alpha_3^{(0)}$	0.318	0.33	0.0172	1
$\alpha_4^{(0)}$	0.636	0.666	0.0242	1
$\alpha_5^{(0)}$	0.953	0.987	0.0308	1
Correlation Parameters				
$\rho^{(1)}$	0.9	0.863	0.0143	1
$\rho^{(0)}$	0.7	0.668	0.0314	1
$\rho^{(10)}$	0.6	0.635	0.0665	1
Mean Treatment Effects[a]				
ATE	0.136	0.149	0.0131	1
TT	0.102	0.113	0.0166	1
LATE	0.141	0.152	0.0193	1
Probabilities of Positive Treatment Effects				
$\Pr(y^{(1)} > y^{(0)})$	0.428	0.452	0.0164	1
$\Pr(y^{(1)} > y^{(0)}\|D(w) = 1)$	0.439	0.461	0.0283	1
$\Pr(y^{(1)} > y^{(0)}\|D(w) = 1, D(\tilde{w}) = 0)$	0.535	0.552	0.0236	1
Probabilities of Zero Treatment Effects				
$\Pr(y^{(1)} = y^{(0)})$	0.42	0.418	0.0248	1
$\Pr(y^{(1)} = y^{(0)}\|D(w) = 1)$	0.492	0.479	0.0272	1
$\Pr(y^{(1)} = y^{(0)}\|D(w) = 1, D(\tilde{w}) = 0)$	0.364	0.363	0.0253	1
Probabilities of Negative Treatment Effects				
$\Pr(y^{(1)} < y^{(0)})$	0.152	0.13	0.0171	1
$\Pr(y^{(1)} < y^{(0)}\|D(w) = 1)$	0.0694	0.0601	0.0145	1
$\Pr(y^{(1)} < y^{(0)}\|D(w) = 1, D(\tilde{w}) = 0)$	0.102	0.0856	0.0204	1

[a] The mean treatment effects are those defined in Section 4 and capture the treatment impacts on the probability that the outcome variable exceeds the lowest category.