CHAPTER 1

BAYESIAN ANALYSIS OF TREATMENT EFFECT MODELS

MINGLIANG LI [1] AND JUSTIN L. TOBIAS[2]

[1]SUNY-Buffalo, USA.
[2]Purdue University, USA.

This chapter reviews Bayesian approaches and Markov chain Monte Carlo (MCMC) methods for estimating treatment-response models. We begin by reviewing the standard continuous outcome / continuous treatment specification under normality and then move on to discuss procedures for handling limited dependent treatment variables and outcomes within this framework. We also discuss methods for relaxing the standard "Gaussian" assumptions commonly made in textbook treatments of this class of problems and commonly seen in empirical applications. In so doing, we discuss issues of model comparison in finite mixture models and conclude with references to some recent work on this topic, including instrument imperfection.

1.1 Introduction

In many fields in Economics - labor, public, industrial organization, international trade and development, to name but a few - the identification and estimation of *causal effects* or *treatment effects* from observational data is a central issue. In these instances the researcher is primarily interested in estimating the effect of some treatment variable, say x, on an outcome, say y, but is concerned that a variety of unobserved factors (i.e., *unobserved confounders*) may be responsible for much of the observed correlation or partial correlation between these variables. To provide a specific and well-studied example, simple regressions of hourly wages (or their logarithm) on education may not provide a good estimate of the causal effect of education on hourly earnings since unobserved ability, motivation, and personality might simultaneously influence both educational attainment and earnings in the labor market.

The dominant technique in the applied literature for estimating causal effects with observational data is to employ *instrumental variables* (IV) or *two-stage least squares (2SLS).* When such methods are employed, the researcher identifies a set of variables which have a conditional impact on the treatment variable x but, given x, are believed to have no other direct influence on the conditional mean of y. Such variables, or *instruments*, serve to both identify the model and can be exploited to consistently estimate the model parameters.

The IV exercise in practice is both an art and a science - conditioned on a valid instrument, or a set of them, one can derive the asymptotic distribution of the IV estimator, test for overidentifying restrictions, investigate inferential consequences when instruments are weak - the usual fare of econometricians and statisticians. Coming up with compelling instruments in actual applied work, however, involves a great deal of imagination and creative skill on the part of the researcher, and the success of such studies depends in no small part on that researcher's ability to form a compelling argument that the instrument in question is plausibly excludable.

While IV / 2SLS and GMM in nonlinear settings remain the most common approaches in literature, a comparably small but growing Bayesian literature also exists for the estimation of treatment-response models and identification of causal effects. It is our goal in this chapter to review some of these methods in a variety of different settings and to discuss estimation of the model parameters via Markov chain Monte Carlo (MCMC) methods.

The outline of this paper is as follows. The following section begins with a standard Gaussian linear treatment-response model and discusses identification and MCMC implementation in that context. Section 1.3 extends these ideas to nonlinear settings. Here we spend some time discussing the popular Roy [57], [31] model and present a general treatment that nests a variety of nonlinear treatment-response specifications. Section 1.4 discusses several departures from these cases, including taking up the case of non-Gaussian errors, model comparison and selection, and issues of instrument imperfection. An illustrative application is presented in Section 1.5, and the chapter concludes with a summary in Section 1.6.

1.2 Linear Treatment Response Models Under Normality

To fix ideas, we begin by considering the following treatment-response model for a continuous outcome y and continuous treatment variable x (throughout using the convention of boldface to denote vectors and bold capitals to denote matrices):

$$\begin{aligned} y_i &= \beta_{y,0} + x_i\beta_{y,x} + \boldsymbol{w}_i\boldsymbol{\beta}_{y,w} + \epsilon_i && (1.1)\\ x_i &= \beta_{x,0} + \boldsymbol{z}_i\boldsymbol{\beta}_{x,z} + u_i, && (1.2)\end{aligned}$$

where

$$\begin{bmatrix}\epsilon_i\\ u_i\end{bmatrix}\Bigg|\boldsymbol{W},\boldsymbol{Z} \overset{iid}{\sim} \mathcal{N}\left[\begin{pmatrix}0\\0\end{pmatrix},\begin{pmatrix}\sigma_\epsilon^2 & \sigma_{\epsilon u}\\ \sigma_{\epsilon u} & \sigma_u^2\end{pmatrix}\right] \equiv \mathcal{N}(\mathbf{0},\boldsymbol{\Sigma}), \quad i = 1, 2, \ldots, n. \qquad (1.3)$$

The system in 1.1 and 1.2, and the type of model we consider throughout this chapter, is *triangular* - meaning that x enters as a right-hand side variable in 1.1, yet y does not enter as a right-hand side variable in 1.2. Thus, the structure here seems appropriate for, say, joint modeling of post-schooling wages (y) and schooling (x), but perhaps not for something like health and income, since in the latter case, we might more naturally consider a fully simultaneous system where health directly affects income, and, conversely, income directly affects health. We also take up the case where there is a single endogenous variable x in 1.1, although the case of several endogenous variables represents a straightforward extension.

Equation 1.3 assumes that the errors are iid bivariate normal. This is a common assumption, and serves as a useful starting point, although it may be potentially inappropriate in many applications. We will take up the issue of relaxing this assumption in Section 1.4. The primary parameter of interest in virtually all cases is the causal effect $\beta_{y,x}$, and unobserved confounders that simultaneously associate with u (and thus x) and ϵ (and thus y) are captured through the covariance parameter $\sigma_{\epsilon u}$, which is suspected as to be non-zero. Finally, $\boldsymbol{W}$ and $\boldsymbol{Z}$, both assumed exogenous, are $n \times k_w$ and $n \times k_z$ matrices (respectively) constructed from the $\boldsymbol{w}_i$ and $\boldsymbol{z}_i$ vectors as follows:

$$\boldsymbol{W} \equiv \begin{bmatrix}\boldsymbol{w}_1\\ \boldsymbol{w}_2\\ \vdots\\ \boldsymbol{w}_n\end{bmatrix}, \quad \boldsymbol{Z} \equiv \begin{bmatrix}\boldsymbol{z}_1\\ \boldsymbol{z}_2\\ \vdots\\ \boldsymbol{z}_n\end{bmatrix}.$$

In empirical work, the elements of $\boldsymbol{w}_i$ and $\boldsymbol{x}_i$ will likely have some degree of overlap, meaning that many elements of $\boldsymbol{w}$ will also be contained in $\boldsymbol{z}$. For example, if y represents earnings and x represents educational attainment, we might believe that variables such as gender, race and ethnicity, test scores, family characteristics, etc., should play a role in both equations. As shown below, however, we will require the appearance of at least one column (or variable) in $\boldsymbol{Z}$ that is not contained in $\boldsymbol{W}$ - these will be our *instruments* which will serve to identify the model parameters and provide a means for parameter estimation.

1.2.1 Instruments and Identification

To see why such an exclusion restriction (or instrument) is necessary in the absence of any additional model structure, let us begin by letting $\boldsymbol{\theta}$ denote all the parameters of the model. We can decompose the bivariate error distribution into the product of a conditional times a marginal:

$$p(\epsilon_i, u_i|\boldsymbol{\theta}) = p(\epsilon_i|u_i, \boldsymbol{\theta})p(u_i|\boldsymbol{\theta}). \tag{1.4}$$

Noting that the Jacobian of the transformation from (ϵ_i, u_i) to (y_i, x_i) is unity (given the triangularity of the model), we obtain

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi(y_i|\mu_{y_i|x}, \sigma^2_{y|x}) \\ &\quad \times\phi(x_i|\beta_{x,0} + \boldsymbol{z}_i\boldsymbol{\beta}_{x,z}, \sigma^2_u), \end{aligned} \tag{1.5}$$

where

$$\begin{aligned} \mu_{y_i|x} &\equiv \beta_{y,0} + x_i\beta_{y,x} + \boldsymbol{w}_i\boldsymbol{\beta}_{y,w} + \frac{\sigma_{\epsilon u}}{\sigma^2_u}(x_i - \beta_{x,0} - \boldsymbol{z}_i\boldsymbol{\beta}_{x,z}) && (1.6) \\ \sigma^2_{y|x} &\equiv \sigma^2_\epsilon(1 - \rho^2_{\epsilon u}) && (1.7) \\ \rho_{\epsilon u} &\equiv \sigma_{\epsilon u}/[\sigma_\epsilon \sigma_u] && (1.8) \end{aligned}$$

and $\phi(s|\mu_s, \sigma^2_s)$ is simply the notation denoting a normal density function for the random variable s with mean μ_s and variance σ^2_s.

It is useful to pause and discuss identification in the context of this system of equations. To this end, let us first consider the case where the set of exogenous covariates are exactly common to both equations in the sense that $\boldsymbol{z}_i = \boldsymbol{w}_i$. In this instance, 1.5 can be written as:

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi(y_i|\psi_0 + x_i\psi_1 + \boldsymbol{w}_i\boldsymbol{\psi}_2, \sigma^2_{y|x}) \\ &\quad \times\phi(x_i|\beta_{x,0} + \boldsymbol{z}_i\boldsymbol{\beta}_{x,z}, \sigma^2_u) \end{aligned} \tag{1.9}$$

where

$$\psi_0 = [\beta_{y,0} - \beta_{x,0}\frac{\sigma_{\epsilon u}}{\sigma^2_u}], \quad \psi_1 = [\beta_{y,x} + \frac{\sigma_{\epsilon u}}{\sigma^2_u}], \quad \boldsymbol{\psi}_2 = [\boldsymbol{\beta}_{y,w} - \boldsymbol{\beta}_{x,z}\frac{\sigma_{\epsilon u}}{\sigma^2_u}]. \tag{1.10}$$

Some quick accounting, then, shows that the likelihood is completely determined by 7 (blocks of) parameters:

$$\beta_{x,0},\ \boldsymbol{\beta}_{x,z},\ \sigma^2_u,\ \psi_0,\ \psi_1,\ \boldsymbol{\psi}_2 \ \text{ and } \ \sigma^2_{y|x} \tag{1.11}$$

whereas we seek to recover 8 "structural" parameters in $\boldsymbol{\theta}$:

$$\beta_{y,0}, \quad \beta_{y,x}, \quad \boldsymbol{\beta}_{y,w}, \quad \beta_{x,0}, \quad \boldsymbol{\beta}_{x,z}, \quad \sigma_u^2, \quad \sigma_\epsilon^2, \quad \text{and} \quad \sigma_{\epsilon u}. \tag{1.12}$$

To summarize, the quantities in 1.11 are identified by the likelihood and consistently estimable whereas the full set of structural parameters in 1.12 is not identifiable. Importantly, observe that the "causal effect" $\beta_{y,x}$ - the object that garners most attention in practice - is among the parameters that are not identifiable when the set of covariates appearing in 1.1, and 1.2 are the same. What is identifiable in this case is $\psi_1 = \beta_{y,x} + \sigma_{\epsilon u}\sigma_u^{-2}$, the "total partial effect" of x on y, which combines the desired causal effect $\beta_{y,x}$ with an effect arising from unobserved confounding.

The mapping between the identified quantities and the structural parameters of interest also reveals the important role that instruments can play in parameter identification. To see this, consider the case where there is at least one variable in $\boldsymbol{z}_i$ that is not contained in $\boldsymbol{w}_i$, say the j^{th} element of $\boldsymbol{z}_i$, or z_i^j. Then observe that the j^{th} element of $\boldsymbol{\beta}_{x,z}$, or $\beta_{x,z}^j$, is clearly identified from the reduced form linear regression in 1.2, and 1.6 reveals that the entire term $\beta_{x,z}^j \sigma_{\epsilon u}/\sigma_u^2$ can be identified as a parameter in the conditional mean $\mu_{y_i|x}$. It follows that the coefficient on the unique element of $\boldsymbol{z}$ that is not contained in $\boldsymbol{w}$ enables us to identify the term arising from unobserved confounding, $\sigma_{\epsilon u}/\sigma_u^2$. Once this ratio is identified, all model parameters, including the causal effect $\beta_{y,x}$, can be recovered.

1.2.1.1 Posterior Simulation As discussed in the introduction to this section, the most common and familiar approaches to estimating models like those in 1.1 and 1.2 employ IV and 2SLS methods. Given a description of the joint distribution as in 1.3, however, a full likelihood-based analysis or Bayesian analysis can also be conducted.

Below we therefore turn to discuss Bayesian estimation of the model parameters in 1.1, 1.2 and 1.3. We do so by employing the *Gibbs sampler* - an iterative simulation method that successively draws from the complete conditional posterior distributions of the model parameters. We do not review the details of Gibbs or MCMC algorithms here, as they are extensively covered elsewhere in this volume.

To begin, let us first stack the variables and parameters as follows:

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} = \begin{bmatrix} 1 & x_i & \boldsymbol{w}_i & 0 & \boldsymbol{0} \\ 0 & 0 & \boldsymbol{0} & 1 & \boldsymbol{z}_i \end{bmatrix} \begin{bmatrix} \beta_{y,0} \\ \beta_{y,x} \\ \boldsymbol{\beta}_{y,w} \\ \beta_{x,0} \\ \boldsymbol{\beta}_{x,z} \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \tag{1.13}$$

or succinctly,

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_i, \tag{1.14}$$

with $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\epsilon}}_i$ defined in the obvious ways. Once stacked in this form, the parameters of this model consist of the vector $\boldsymbol{\beta}$ and the elements of the covariance matrix $\boldsymbol{\Sigma}$. A Gibbs sampling algorithm, then, will require us to derive and sample from the posterior conditionals $\boldsymbol{\beta}|\boldsymbol{\Sigma}, \tilde{\boldsymbol{y}}$ and $\boldsymbol{\Sigma}|\boldsymbol{\beta}, \tilde{\boldsymbol{y}}$. To this end, suppose we are to employ priors of the forms:

$$\begin{aligned} \boldsymbol{\beta} &\sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta}) \qquad (1.15) \\ \boldsymbol{\Sigma}^{-1} &\sim W\left([\kappa \boldsymbol{R}]^{-1}, \kappa\right), \qquad (1.16) \end{aligned}$$

with $\mathcal{N}(\cdot,\cdot)$ denoting a multivariate normal distribution with the given mean and variance and $W(\cdot,\cdot)$ denoting a Wishart distribution, parameterized as in Koop, Poirier and Tobias [41], page 339.

With this done, posterior simulation in our linear model with an endogeneity problem follows in a straightforward way. Specifically, a simple two-block Gibbs algorithm can be employed that iteratively samples from

$$\boldsymbol{\beta}|\boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{D_\beta d_\beta}, \boldsymbol{D_\beta}), \qquad (1.17)$$

where

$$\boldsymbol{D_\beta} = \left(\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i\right)^{-1}, \quad \boldsymbol{d_\beta} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\boldsymbol{\mu_\beta} + \sum_{i=1}^{n}\left(\tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{y}}_i\right) \qquad (1.18)$$

and

$$\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{x} \sim W\left(\left[\kappa \boldsymbol{R} + \sum_{i=1}^{n} \tilde{\boldsymbol{\epsilon}}_i \tilde{\boldsymbol{\epsilon}}_i{}'\right]^{-1}, \kappa + n\right). \qquad (1.19)$$

For the interested reader, details of these derivations can be found in Lindley and Smith [48] or Koop, Poirier and Tobias [41].

A posterior simulator for this model proceeds by iteratively sampling from 1.17 and 1.19, always conditioning on the most recent parameters produced from the sampling scheme. Once convergence to the posterior is determined to have been achieved, the subsequent samples can be used to obtain point estimates (e.g. posterior means), quantiles or entire posterior distributions for objects of interest. We provide an illustrative application of how this is done in Section 1.5. It is worth emphasizing, however, that estimation in this linear treatment-response model is nearly as simple as IV or 2SLS: all that is required is the generation of normal and Wishart variates, and the tiniest dose of patience as one waits a few seconds for the software to return a suitable post-convergence sample of parameters.

We close this section noting that the reader somewhat familiar with Bayes might observe, and perhaps be puzzled by, the connection of the simulator in 1.17 and 1.19 to the Gibbs algorithm one would obtain from a standard system of equations analysis such as a seemingly unrelated regressions (SUR) model. That is, one might find himself or herself asking: why does the simulator for this model with an endogeneity problem reduce to essentially the same simulator that would be used to estimate a bivariate SUR without any endogeneity concerns? The connection here critically relies on the Jacobian of transformation from (ϵ, u) to (x, y) being equal to one; such a result would not be obtained for a purely simultaneous equations model that is not triangular (e.g., the fully simultaneous income / health example mentioned previously).

1.3 Nonlinear Treatment Response Models

The previous section offered a model and estimation procedure for assessing the causal effect of treatment when both the outcome y and treatment variable x were continuous. In many (possibly most) cases, however, at least one of these variables will be discrete in nature. To provide just a few observational data examples, the treatment x might indicate participation in a job training program, or high school graduation. In each of these cases, some modification of the system in 1.1 and 1.2 is called for in order to properly account for the binary nature of the treatment.

If the outcome y in these cases (like a log wage) remains continuous, while the treatment x is binary, an extension of 1.2 and 1.1 would be to consider:

$$\begin{aligned} y_i &= \beta_{y,0} + x_i\beta_{y,x} + \boldsymbol{w}_i\boldsymbol{\beta}_{y,w} + \epsilon_i && (1.20)\\ x_i^* &= \beta_{x,0} + \boldsymbol{z}_i\boldsymbol{\beta}_{x,z} + u_i, && (1.21)\\ x_i &= I(x_i^* > 0). && (1.22) \end{aligned}$$

A key feature of the model in 1.20 and 1.21 is the addition of *latent data* (in this case x_i^*), and the adoption of a model that, like 1.1 and 1.2, is linear, but now is represented as linear in the latent-data x_i^* as opposed to the observed outcome x_i . In the above example one might interpret x_i^* as the (unobserved by the econometrician) net desire for receipt of treatment. The observed binary outcome x_i takes the value of one if this net desire is positive, and otherwise equals zero. This mapping between the latent construct x_i^* and the observed outcome x_i is formalized in 1.22 - the agent takes the treatment if her net desire for doing so is positive, and otherwise is left untreated.

The system in 1.20, 1.21 and 1.22 offers just one example of a nonlinear treatment-response system where y is continuous and x is binary. There are a variety of other cases, however, for us to consider - for example, the outcome y_i may be binary rather than continuous, leading us to consider a latent-variable version of 1.20 and adding another link between observed and latent outcomes, as in 1.22. Other cases that arise commonly in empirical work include *ordinal* outcomes. Below we consider a fairly generic nonlinear treatment-response system, applicable to a variety of data types.

1.3.1 A General Nonlinear Representation

We begin by defining the vectors: $\tilde{\boldsymbol{y}}_i^* = [y_i^* \; x_i^*]'$ and $\tilde{\boldsymbol{y}}_i = [y_i \; x_i]'$, which represent a 2×1 latent data vector and 2×1 observed data vector, respectively.

With these definitions in place, a reasonably general nonlinear treatment-response system can be described as follows:

$$\begin{aligned} \tilde{\boldsymbol{y}}_i^* | \tilde{\boldsymbol{X}}, \boldsymbol{\beta} &\overset{ind}{\sim} \mathcal{N}\left(\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}, \boldsymbol{\Sigma}\right) &(1.23)\\ \tilde{\boldsymbol{y}}_i | \tilde{\boldsymbol{y}}_i^* &= g(\tilde{\boldsymbol{y}}_i^*, \boldsymbol{\alpha}), \quad i = 1, 2, \ldots, n. &(1.24) \end{aligned}$$

Equation 1.23 represents a latent-variable version of 1.14. Equation 1.24 then links the latent data to the observed outcome pair $\tilde{\boldsymbol{y}}_i$ through the function $g(\cdot)$ and, potentially, a vector of parameters $\boldsymbol{\alpha}$. For example, if both the outcome y_i and treatment x_i are binary, 1.24 would become:

$$y_i = I(y_i^* > 0), \quad x_i = I(x_i^* > 0).$$

If y_i was instead a positive variable censored at zero (such as expenditure or wages, for example) while x_i remained binary, we could write:

$$y_i = \max\{0, y_i^*\}, \quad x_i = I(x_i^* > 0).$$

As a final example, if both x_i and y_i were ordinal responses, we could specify

$$\begin{aligned} y_i &= j \text{ if } \alpha_j^{(y)} < y_i^* \leq \alpha_{j+1}^{(y)}, \quad j = 1, 2, \ldots, J_y \\ x_i &= l \text{ if } \alpha_l^{(x)} < x_i^* \leq \alpha_{l+1}^{(x)}, \quad l = 1, 2, \ldots, L_x. \end{aligned}$$

In this final case, unlike the binary and censored cases, the link function g in 1.24 depends on a vector of *cutpoint* parameters $\boldsymbol{\alpha}$. In specifications where these parameters are present, additional steps will be added to the posterior simulator in order to generate $\boldsymbol{\alpha}$ samples.

1.3.1.1 Gibbs Implementation Let $\boldsymbol{\theta} = [\boldsymbol{\beta}' \; vec(\boldsymbol{\Sigma})' \; \boldsymbol{\alpha}']'$ denote all the parameters in this specification. We will assume prior independence among these components, continue to use a multivariate normal prior for $\boldsymbol{\beta}$: $\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu_\beta}, \boldsymbol{V_\beta})$, as in 1.15 and a Wishart prior for $\boldsymbol{\Sigma^{-1}}$: $\boldsymbol{\Sigma^{-1}} \sim W([\kappa \boldsymbol{R}]^{-1}, \kappa)$ as in 1.16 and leave the prior for $\boldsymbol{\alpha}$ generically specified as $p(\boldsymbol{\alpha})$.

In order to implement a Gibbs algorithm for this model, we need to derive and sample from the complete conditional posterior distributions of the model parameters. If the $\tilde{\boldsymbol{y}}_i^*$ were "known," posterior inference could proceed very similarly to the continuous outcome model presented in 1.1 and 1.2. That is, $\boldsymbol{\beta}|\tilde{\boldsymbol{y}}^*, \boldsymbol{\Sigma}$ would be multivariate normal, and $\boldsymbol{\Sigma}^{-1}|\boldsymbol{\beta}, \tilde{\boldsymbol{y}}^*$ would remain a Wishart distribution, although possibly with some restrictions on its elements.

Given these appealing conveniences afforded by conditioning on $\tilde{\boldsymbol{y}}^*$, a possible approach is to *augment* the posterior distribution with the latent data $\tilde{\boldsymbol{y}}^*$ and sample the latent data vectors in the course of our posterior simulation [e.g., Tanner and Wong [59] and Albert and Chib [1]]. We therefore consider the augmented posterior $p(\boldsymbol{\theta}, \tilde{\boldsymbol{y}}^*|\tilde{\boldsymbol{y}})$. A Gibbs implementation, then, requires us to derive and sample from the four posterior conditionals for $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}^{-1}$, $\tilde{\boldsymbol{y}}^*$ and $\boldsymbol{\alpha}$ (and the last of these only when required).

The first two conditionals, as suggested earlier, are easily sampled conditional on the latent data:

$$\boldsymbol{\beta}|\boldsymbol{\Sigma}, \tilde{\boldsymbol{y}}^*, \tilde{\boldsymbol{y}} \sim \mathcal{N}\left(\boldsymbol{D_\beta d_\beta}, \boldsymbol{D_\beta}\right) \tag{1.25}$$

where

$$\boldsymbol{D_\beta} \equiv \left[\left(\sum_{i=1}^n \tilde{\boldsymbol{X}}_i' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i\right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1}\right]^{-1}, \quad \boldsymbol{d_\beta} \equiv \left(\sum_{i=1}^n \tilde{\boldsymbol{X}}_i' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{y}}_i^*\right) + \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu_\beta}, \tag{1.26}$$

and

$$\boldsymbol{\Sigma^{-1}}|\boldsymbol{\beta}, \tilde{\boldsymbol{y}}^*, \tilde{\boldsymbol{y}} \sim W\left(\left[\kappa \boldsymbol{R} + \sum_{i=1}^n (\tilde{\boldsymbol{y}}_i^* - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})(\tilde{\boldsymbol{y}}_i^* - \tilde{\boldsymbol{X}}_i \boldsymbol{\beta})'\right]^{-1}, \kappa + n\right). \tag{1.27}$$

Note that, in terms of our coding, we act as if the latent data $\tilde{\boldsymbol{y}}^*$ are observed, and will simply update this vector at each step in our sampler. To complete a description of our posterior simulator, we note that the joint posterior distribution for the latent $\tilde{\boldsymbol{y}}^*$ and vector of parameters $\boldsymbol{\alpha}$ is given as:

$$p(\tilde{\boldsymbol{y}}^*, \boldsymbol{\alpha}|\tilde{\boldsymbol{y}}, \boldsymbol{\beta}, \boldsymbol{\Sigma}) \propto p(\boldsymbol{\alpha}) \prod_{i=1}^n \phi(\tilde{\boldsymbol{y}}_i^*|\boldsymbol{X}_i \boldsymbol{\beta}, \boldsymbol{\Sigma}) I\left[\tilde{\boldsymbol{y}}_i = g(\tilde{\boldsymbol{y}}_i^*, \boldsymbol{\alpha})\right]. \tag{1.28}$$

To fix ideas and provide a little clarity to 1.28, consider the case discussed at the outset of this section where y is continuous and x is binary. In this instance, there are no parameters in $\boldsymbol{\alpha}$, and y_i^* is not needed, since that component of the model is fully observed. It remains, then, to sample just the latent data x_i^*. We do so by noting that 1.28 reduces to:

$$\begin{aligned} p(x_i^*|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{x}, \boldsymbol{y}) \quad \propto \quad & \phi(x_i^*|\mu_{x_i^*|y}, \sigma_{x|y}^2) \times \\ & \left[I(x_i^* > 0)I(x_i = 1) + I(x_i^* \leq 0)I(x_i = 0)\right], \end{aligned} \tag{1.29}$$

for $i = 1, 2, \ldots, n$, where

$$\mu_{x_i^*|y} = \beta_{x,0} + \boldsymbol{z}_i \boldsymbol{\beta}_{x,z} + \frac{\sigma_{\epsilon u}}{\sigma_\epsilon^2}\left[y_i - \beta_{y,0} - x_i \beta_{y,x} - \boldsymbol{w}_i \boldsymbol{\beta}_{y,w}\right] \tag{1.30}$$

$$\sigma_{x|y}^2 = \sigma_u^2 (1 - \rho_{u\epsilon}^2). \tag{1.31}$$

The structure of 1.29 clearly reveals that the sampling of each x_i^* can be conducted by generating draws from univariate truncated normal distributions. Simulations from the truncated normal are easily obtained, and can be produced directly via the method of inversion (see, e.g., Koop, Poirier and Tobias [41], exercise 11.20).

When $\boldsymbol{\alpha}$ remains a component of the posterior simulator, owing to the consideration of, for example, ordinal data, an additional sampling step is required. While one could potentially proceed by sampling from $\{\tilde{\boldsymbol{y}}_i^*|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{y}}\}_{i=1}^n$

and $\boldsymbol{\alpha}|\tilde{\boldsymbol{y}}^*, \boldsymbol{\beta}, \boldsymbol{\Sigma}, \tilde{\boldsymbol{y}}$, this is usually not advisable given the high degree of correlation between the latent data $\tilde{\boldsymbol{y}}^*$ and the parameters in $\boldsymbol{\alpha}$. In this situation, $\tilde{\boldsymbol{y}}^*$ can often be integrated out of 1.28, $\boldsymbol{\alpha}$ can then be drawn from the resulting marginalized conditional, and finally the latent data can be drawn independently from its complete conditional:

$$p(\tilde{\boldsymbol{y}}_i^*|\boldsymbol{\beta}, \boldsymbol{\Sigma}, \boldsymbol{\alpha}, \tilde{\boldsymbol{y}}) \quad \propto \quad \phi(\tilde{\boldsymbol{y}}_i^*|\boldsymbol{X}_i\boldsymbol{\beta}, \boldsymbol{\Sigma})I[\tilde{\boldsymbol{y}}_i = g(\tilde{\boldsymbol{y}}_i^*, \boldsymbol{\alpha})], \quad i = 1, 2, \ldots, n,$$

which typically amounts, again, to truncated normal sampling. Jeliazkov et al. [35] discuss a number of important issues related to identification of univariate and multivariate ordinal models and evaluate the performances of alternate schemes for sampling $(\boldsymbol{\alpha}, \tilde{\boldsymbol{y}}^*)$ in such models.

An important issue that, to this point, has only been briefly and casually referenced, concerns restrictions on elements of the covariance matrix $\boldsymbol{\Sigma}$. For example, in the case of binary observed outcomes, a common identification restriction is to impose that the diagonal elements of $\boldsymbol{\Sigma}$ are unity. In this case, the sampling of $\boldsymbol{\Sigma}^{-1}$ is not from 1.27, but rather, from 1.27 with restrictions on the main diagonal.

A few alternatives exist for dealing with this problem. Reparameterizations (e.g., Li [43] and McCulloch, Polson and Rossi [49]) are sometimes possible which allow the researcher to essentially sidestep the restrictions, sampling elements of $\boldsymbol{\Sigma}$ in blocks while preserving the positive definiteness of the covariance matrix. In the case of a single diagonal restriction, as would be the case if only one of the outcomes were binary, Nobile [54] comments on the reparameterization scheme of McCulloch, Polson and Rossi [49], and provides a way to directly sample from a Wishart, given a restriction on a diagonal element. Chan and Jeliazkov [4] provide additional details regarding sampling from restricted covariance matrices. Finally, and consistent with our presentation in this chapter, one can choose to simply ignore the restrictions that are in place when implementing the simulator and post-process the simulations to focus on identified quantities. That is, one can simply sample $\boldsymbol{\Sigma}^{-1}$ as in 1.27 and appropriately adjust those simulations at each iteration of the sampler to calculate quantities that are identifiable. Rossi, Allenby and McCulloch [56], particularly in chapters 3 and 4, discuss applications of this approach and compare performances of samplers that navigate through identified and non-identified parameter spaces.

1.3.1.2 The Roy Model A popular specification in treatment-response modeling, and a slight generalization of the system in 1.20 and 1.21, is the Roy [57] model or model of *potential outcomes* (see also Heckman and Honoré [31]). This specification again considers the case of a binary treatment, (and as such 1.21 and 1.22 remain unchanged), but adds to 1.20 by explicitly modeling outcome equations for both the treated and untreated regimes. Specifically, we might represent a slightly generalized version of the Roy model as:

$$\begin{aligned} y_i^{(1)} &= \beta_{y,0}^{(1)} + \boldsymbol{w}_i \boldsymbol{\beta}_{y,w}^{(1)} + \epsilon_i^{(1)} \qquad (1.32) \\ y_i^{(0)} &= \beta_{y,0}^{(0)} + \boldsymbol{w}_i \boldsymbol{\beta}_{y,w}^{(0)} + \epsilon_i^{(0)} \qquad (1.33) \\ x_i^* &= \beta_{x,0} + \boldsymbol{z}_i \boldsymbol{\beta}_{x,z} + u_i. \qquad (1.34) \end{aligned}$$

Variables and parameters associated with the treated state have been assigned a superscript of (1) in 1.32, while those associated with the untreated state are assigned a superscript of (0) in 1.33. Treatment effects are summarized by the outcome gain (or loss) resulting from receipt of treatment, denoted as $y^{(1)} - y^{(0)}$, and trivariate normality among $[\epsilon_i^{(1)} \ \epsilon_i^{(0)} \ u_i]'$ is assumed as a starting point, with the methods described in the following section used to relax this assumption, when needed.

In terms of posterior simulation, simple generalizations of the sampling steps used for the estimation of 1.20 and 1.21 can be employed to fit this model. Specifically, given the complete data, the posterior conditional for the stacked vector of parameters $\boldsymbol{\beta}$ will again remain normal with a structure identical to 1.25, while that for $\boldsymbol{\Sigma}^{-1}$ will remain Wishart, as in 1.27, with a diagonal restriction that sets $\sigma_u^2 = 1$. As discussed previously, this sampling step for $\boldsymbol{\Sigma}^{-1}$ can either be performed via a reparameterization, or a draw from a restricted Wishart can be directly obtained (e.g., Nobile [54]), or the restriction can be ignored and the simulations post-processed to focus on identifiable quantities. As for the sampling of the latent data, the x_i^* are again sampled independently from their conditional truncated normal distributions, readily derived from the trivariate system of equations.

We also observe that for each observation i, either the treated or the untreated outcome is observed for each agent (but never both), and as such the observed outcome y_i can be expressed as

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

Thus, the treatment effect $\Delta \equiv y^{(1)} - y^{(0)}$ involves a *counterfactual* - the outcome the agent would have received had he or she made a different decision regarding the receipt of treatment. For every individual, then, exactly one of the $y^{(j)}, \ j = 0, 1$, is missing while the other is observed. In the spirit of data augmentation, we can thus add this missing outcome as an element of the joint posterior, and in the course of implementing the posterior simulator, sample the"missing" outcome for each i from its conditional normal distribution. Conditioned on this missing data - which is updated at every iteration of the sampler - the sampling of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ proceeds as if we were faced with a fully observed system of linear equations (e.g., a seemingly unrelated regressions model). Full details regarding a Gibbs algorithm for this model are provided in Chib and Hamilton [11] and Koop, Poirier and Tobias [41], exercise 14.11, and thus are not repeated here.

A final interesting issue arises in the Roy model setting, and is related to the potential outcomes structure it embraces. Since only $y^{(1)}$ *or* $y^{(0)}$ is observed for each individual, the correlation parameter between the errors of the treated and untreated equations, which we could denote as ρ_{10} to fix ideas, does not enter the likelihood function for the observed data and thus is not identified. Despite this, Vijverberg [60] notes the possibility of learning about ρ_{10}, an idea formalized in the Bayesian setting by Koop and Poirier [39]. These authors clearly illustrate that the marginal priors and posteriors for ρ_{10} will not be the same, in general, even though this parameter is not identified by the likelihood function. The vehicle for this updating of marginal prior beliefs is prior dependence - information learned about identified correlation parameters, coupled with restrictions that $\mathbf{\Sigma}$ must be positive semidefinite, creates a knowledge spillover which restricts the conditional support of ρ_{10}. Chib and Hamilton [11] address the non-identification of the cross-regime correlation ρ_{10} by setting it equal to zero and fitting the model subject to that restriction, while Poirier and Tobias [55] and Li, Poirier and Tobias [47] allow for learning about this parameter and provide several applications. In a recent statement on this issue, Chib [8] recommends working directly in the identified model space as opposed to the augmented potential outcomes space, as doing so improves the mixing properties of the posterior simulator and frees the researcher of the need to worry about the influence of the conditional prior for ρ_{10}- whose influence on posterior results does not vanish even asymptotically.

1.4 Other Issues and Extensions: Non-Normality, Model Selection and Instrument Imperfection

1.4.1 Non-Normality

The reader may have observed that all the analysis to this point has been conducted based on the assumption of joint normality of the error terms. While the assumption of normality does serve as a useful starting point, and may suffice as an adequate description of the data in some cases, it is necessary nonetheless to have a modeling framework allowing for significantly greater flexibility.

When seeking out these more general treatments, it seems prudent to keep in mind a desire to keep things computationally tractable. What has been evident from our previous examples is that normal likelihoods combine nicely with conditionally conjugate normal / Wishart priors to yield conditional posterior distributions that are easily sampled, thus facilitating Gibbs estimation. These computational conveniences of normality are something we would like to retain when considering more general specifications to take to the data.

One possibility, as noted by Carlin and Polson [3] and Geweke [24], and illustrated for the case of treatment-response modeling by Chib and Hamilton [11], is to consider a class of Student-t sampling models by representing them as a scale mixture of multivariate Gaussian distributions. Specifically, one can generalize 1.23 by assuming

$$\begin{aligned}
\tilde{\boldsymbol{y}}_i^* &\overset{ind}{\sim} \mathcal{N}(\tilde{\boldsymbol{X}}_i\boldsymbol{\beta}, \lambda_i\boldsymbol{\Sigma}) && (1.35)\\
\lambda_i &\overset{iid}{\sim} IG\left(\frac{\nu}{2}, \frac{2}{\nu}\right) \quad \Rightarrow p(\lambda_i) \propto \lambda_i^{-([\nu/2]+1)} \exp\left(-\frac{\nu}{2\lambda_i}\right), && (1.36)
\end{aligned}$$

where IG denotes an inverse gamma distribution, whose density function is provided up to proportionality.

When 1.35 is marginalized over the scale mixing variables λ_i, whose prior is given in 1.36, a multivariate-t sampling model (with ν degrees of freedom) for $\tilde{\boldsymbol{y}}_i^*$ is produced. This extension thus allows the researcher to address issues of fat tails beyond those allowed by the normal distribution, and the conditional normal representation of the sampling model leads to tractable conditional posteriors amenable to Gibbs sampling.

Gibbs implementation in the Student-t sampling model is fully described in Chib and Hamilton [11] and we briefly overview those details here. Conditioned on $\boldsymbol{\lambda} = [\lambda_1, \lambda_2, \cdots, \lambda_n]'$, the sampling of $\tilde{\boldsymbol{y}}^*$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$ follow similarly to what was described in the previous section. Specifically, the sampling of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}$ will involve the simulation of normal and (possibly restricted) inverse Wishart variates, respectively, while the sampling of $\tilde{\boldsymbol{y}}^*$ in many cases will involve truncated normal sampling. The posterior conditional for each λ_i can be shown to be of the inverse gamma form, and draws from the inverse gamma can simply be obtained by inverting gamma random variates. Thus, extension to the Student-t case just adds an additional step to the Gibbs algorithm for the standard Gaussian model, and the conditional normal scale mixture representation of the Student-t in 1.35 and 1.36 retains the computational conveniences afforded by normality.

A more general representation that also allows for possible skewness or multimodality in the error distribution is to consider a finite mixture sampling model for the augmented data:

$$\begin{aligned}
\tilde{\boldsymbol{y}}_i^*|c_i &\sim \mathcal{N}\left(\boldsymbol{\beta}_{0c_i} + \tilde{\boldsymbol{X}}_i\tilde{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{c_i}\right), && (1.37)\\
\Pr(c_i = g|\boldsymbol{\pi}) &= \pi_g, \quad g = 1, 2, \ldots, G, \quad \sum_{g=1}^{G} \pi_g = 1, \pi_g > 0 \ \forall g, && (1.38)
\end{aligned}$$

where $\boldsymbol{\beta}_{0c_i} = [\beta_{y,0c_i}, \beta_{x,0c_i}]'$ denotes component specific intercepts and $\tilde{\boldsymbol{\beta}}$ represents slope coefficients that are common to all components. The variable c_i is an integer-valued component labeling variable with support over $g = 1, 2, \cdots, G$, representing which mixture component observation i is generated from. The vector $\boldsymbol{\pi} = [\pi_1, \pi_2, \cdots, \pi_G]'$ is a vector of component probabilities, and marginalized over c_i, the sampling model for $\tilde{\boldsymbol{y}}_i^*$ is easily seen as a mixture (or average) of multivariate normal models. To complete the model, a Dirichlet prior for $\boldsymbol{\pi}$ is commonly assumed.

In terms of posterior simulation, the sampling of $\boldsymbol{\beta} \equiv [\boldsymbol{\beta}_{01}', \boldsymbol{\beta}_{02}', \cdots, \boldsymbol{\beta}_{0G}', \tilde{\boldsymbol{\beta}}']'$ and $\boldsymbol{\Sigma}^{-1} \equiv [\boldsymbol{\Sigma}_1^{-1}, \boldsymbol{\Sigma}_2^{-1}, \cdots, \boldsymbol{\Sigma}_G^{-1}]$, follow essentially the same as in 1.25 and

1.27, although only the subset of observations currently assigned to the g^{th} component are used in the respective inference for the component specific intercepts and covariance matrices. The component labeling variables themselves are drawn from a discrete distribution with support over $g = 1, 2, \cdots, G$, and the component probabilities are sampled from a Dirichlet. Details of these calculations are provided in Chib and Hamilton [11] and Li, Poirier and Tobias [47], and are not repeated here for the sake of brevity.

Finite mixture models are quite flexible and can adapt to capture a variety of features of the data distribution, including skewness and multimodality. In some cases, prior knowledge of the problem at hand suggests that the population of interest consists of distinct groups, whose responses are likely to be homogeneous within each group but potentially different across groups. In this case it is natural to adopt a finite mixture model and to interpret its parameters. In other (perhaps most) cases, finite mixtures are simply used as a flexible modeling device to allow for departures from normality, and any ex-post interpretation of results as evidence for the existence of discrete "groups" in the population should be made with caution.

In practice, use of finite mixtures can produce concerns regarding parameter and component identification (i.e., the model parameters are not identified up to a relabeling of the components and the simulator can exhibit label-switching). Furthermore, the choice of the number of mixture components is a question that should be addressed, and we take up this issue in the following section.

Finally, an even more general approach, as described in Conley, Hansen, McCulloch and Rossi [17], is to consider an analysis based on Dirichlet process (DP) priors. In this approach, parameters are observation-specific, and assumed to be generated from a common distribution. A prior is placed over this common distribution, centered around an overall base measure. Although this approach avoids explicit selection of the number of components, one can think of this as being handled endogenously, as new parameter values (quite similar in spirit to components) are "born" when needed, "die" when unneeded, and are lumped together when appropriate to retain parsimony. Neal [53] provides a number of useful details related to posterior simulation in DP models.

1.4.2 Model Comparison

As discussed in the previous section, when finite mixture models are adopted, there is a need to determine the appropriate number of mixture components. Chib [7] and Chib and Jeliazkov [16] offer very useful, general-purpose methods for model comparison, model averaging and marginal likelihood calculation in a Bayesian setting. Below we discuss an alternative approach to this problem based on the bridge sampling technique.

A special case of the bridge sampling method is the reciprocal importance sampling approach utilized in Gelfand and Dey [23]:

$$p(\tilde{\boldsymbol{y}}) \quad \approx \quad \left[M^{-1}\sum_{m=1}^{M}\frac{q(\boldsymbol{\theta}^{(m)})}{p(\boldsymbol{\theta}^{(m)})p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}^{(m)})}\right]^{-1} \tag{1.39}$$

where $\boldsymbol{\theta}^{(m)}$ is the m^{th} draw of $\boldsymbol{\theta} = [\boldsymbol{\beta}', vec(\boldsymbol{\Sigma})', \boldsymbol{\alpha}', \boldsymbol{\pi}']'$ from a Gibbs algorithm for the model in 1.37 and 1.38. The denominator in the above, $p(\boldsymbol{\theta}^{(m)})p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}^{(m)})$, is an evaluation of the non-normalized posterior distribution of $\boldsymbol{\theta}$, which is the product of the prior distribution $p(\boldsymbol{\theta})$ and the likelihood function $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$ marginalized over the latent outcomes $\tilde{\boldsymbol{y}}^*$ and component labeling variables $\boldsymbol{c} = [c_1, c_2, \cdots, c_n]'$. The numerator in the sampling average $q(\boldsymbol{\theta}^{(m)})$ is an evaluation of the importance density $q(\boldsymbol{\theta})$. In the rare case where the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ has a known exact analytical form with an explicit expression of the normalizing constant, the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ serves as the best candidate for $q(\boldsymbol{\theta})$. In fact, in such a case, it is not even necessary to make the sampling average over the M draws of $\boldsymbol{\theta}^{(m)}$ and 1.39 holds exactly even for $M = 1$. Chib [7] and Chib and Jeliazkov [16] follow this strategy and make use of the Gibb algorithm to produce an accurate evaluation of the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ at a particular draw $\boldsymbol{\theta}^{(m)}$.

Often the importance density $q(\boldsymbol{\theta})$ chosen by a researcher has fatter tails than the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$, making the reciprocal importance sampling method not very reliable. Various researchers, including Geweke [25], Meng and Wong [50] and Frühwirth-Schnatter [22], have proposed important improvements over the reciprocal importance sampling method. For example, the bridge sampling technique examined and applied in Meng and Wong [50] and Frühwirth-Schnatter [22] generalize and replace 1.39 with

$$p(\tilde{\boldsymbol{y}}) \quad \approx \quad \frac{L^{-1}\sum_{l=1}^{L}\alpha(\tilde{\boldsymbol{\theta}}^{(l)})p(\tilde{\boldsymbol{\theta}}^{(l)})p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}^{(l)})}{M^{-1}\sum_{m=1}^{M}\alpha(\boldsymbol{\theta}^{(m)})q(\boldsymbol{\theta}^{(m)})} \tag{1.40}$$

where $\tilde{\boldsymbol{\theta}}^{(l)}$, for $l = 1, 2, \cdots, L$, are L independent draws from the importance density $q(\boldsymbol{\theta})$. Various choices of the function $\alpha(\boldsymbol{\theta})$ result in different versions of the bridge sampling method. For example, the reciprocal importance sampling approach in 1.39 is a special case of the bridge sampling method with $\alpha(\boldsymbol{\theta}) = [p(\boldsymbol{\theta})p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})]^{-1}$.

Meng and Wong [50] propose an asymptotically optimal choice of $\alpha(\boldsymbol{\theta})$ under certain conditions:

$$\alpha(\boldsymbol{\theta}) \quad \propto \quad \frac{1}{Lq(\boldsymbol{\theta}) + Mp(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})}. \tag{1.41}$$

Since the exact form of the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}) = [p(\boldsymbol{\theta})p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})]/p(\tilde{\boldsymbol{y}})$ in turn depends on the knowledge of the marginal likelihood $p(\tilde{\boldsymbol{y}})$, in practice, Meng and Wong [50] suggest using some starting value of $p(\tilde{\boldsymbol{y}})$ such as the output from the reciprocal importance sampling approach in 1.39:

$$p_0(\tilde{\boldsymbol{y}}) = \left[M^{-1} \sum_{m=1}^{M} \frac{q(\boldsymbol{\theta}^{(m)})}{p(\boldsymbol{\theta}^{(m)}) p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta}^{(m)})} \right]^{-1},$$

and then, for $t = 1, 2, \cdots, T$, iteratively updating the values of $p_t(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ and $p_t(\tilde{\boldsymbol{y}})$ using the following two steps which can converge rather quickly to the bridge sampling estimate of $p(\tilde{\boldsymbol{y}})$ with an asymptotically optimal choice of $\alpha(\boldsymbol{\theta})$:

$$p_t(\boldsymbol{\theta}|\tilde{\boldsymbol{y}}) = \frac{p(\boldsymbol{\theta}) p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})}{p_{t-1}(\tilde{\boldsymbol{y}})} \tag{1.42}$$

$$\begin{aligned} p_t(\tilde{\boldsymbol{y}}) = & \left[L^{-1} \sum_{l=1}^{L} \frac{p(\tilde{\boldsymbol{\theta}}^{(l)}) p(\tilde{\boldsymbol{y}}|\tilde{\boldsymbol{\theta}}^{(l)})}{Lq(\tilde{\boldsymbol{\theta}}^{(l)}) + Mp_t(\tilde{\boldsymbol{\theta}}^{(l)}|\tilde{\boldsymbol{y}})} \right] \\ & \times \left[M^{-1} \sum_{m=1}^{M} \frac{q(\boldsymbol{\theta}^{(m)})}{Lq(\boldsymbol{\theta}^{(m)}) + Mp_t(\boldsymbol{\theta}^{(m)}|\tilde{\boldsymbol{y}})} \right]^{-1}. \end{aligned} \tag{1.43}$$

Meng and Wong [50] and Frühwirth-Schnatter [22] show many advantages of the bridge sampling method over the reciprocal importance sampling approach. However, for various finite mixture models, including the one in 1.37 and 1.38, the choice of the importance density $q(\boldsymbol{\theta})$ remains critical. It is well known that the finite mixture models suffer from an identification problem in the sense that both the prior $p(\boldsymbol{\theta})$ and the likelihood function $p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$ in most cases are labeling-invariant. In such cases, the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ which is proportional to $p(\boldsymbol{\theta})p(\tilde{\boldsymbol{y}}|\boldsymbol{\theta})$ also will not change if we relabel all the G components completely. Unless there exist some restrictions, such as $\beta_{x,01} > \beta_{x,02} > \cdots > \beta_{x,0G}$, imposed on the prior $p(\boldsymbol{\theta})$, the likelihood function itself does not distinguish among all the $G!$ different ways of component labeling. Related to this identification problem, a typical Gibbs algorithm for finite mixture modeling also suffers from a slow mixing problem. In theory, the Gibbs sampler should converge to draws from the posterior distribution $p(\boldsymbol{\theta}|\tilde{\boldsymbol{y}})$ by visiting each of the $G!$ labeling subspaces. But in practice, as explained in Frühwirth-Schnatter [21], [22] and Geweke [27], the Gibbs draws tend to stick in only one of the $G!$ labeling subspaces even with a large number of draws.

To combat the above mentioned identification problem in finite mixture modeling directly, Frühwirth-Schnatter [21] proposes a random permutation method to force the Gibbs sampler to visit each of the $G!$ labeling subspaces. For the model in 1.37 and 1.38, this random permutation approach is achieved by, at the end of the m^{th} draw of the Gibbs algorithm, reassigning the values of the parameters in $\boldsymbol{c}^{(m)}$, $\boldsymbol{\pi}^{(m)}$, $\boldsymbol{\beta}^{(m)}$ and $\boldsymbol{\Sigma}^{(m)}$ that are component-specific to one of the $G!$ labeling schemes that is randomly chosen. In this way, each of the

$G!$ subspaces is visited with a probability of exactly $1/G!$. Correspondingly, to compute the marginal likelihood in (1.43), one can choose an importance density following the suggestions in Frühwirth-Schnatter [22] and Kaufmann and Frühwirth-Schnatter [36]:

$$\begin{aligned} q(\boldsymbol{\theta}) &= \frac{1}{S}\sum_{s=1}^{S} p(\boldsymbol{\pi}|\boldsymbol{c}^{(s)})p(\boldsymbol{\beta}|\boldsymbol{c}^{(s)},\boldsymbol{\Sigma}^{(s-1)},\tilde{\boldsymbol{y}}^{*(s-1)}) \qquad (1.44)\\ &\quad \times p(\boldsymbol{\Sigma}|\boldsymbol{c}^{(s)},\boldsymbol{\beta}^{(s)},\tilde{\boldsymbol{y}}^{*(s-1)})q(\boldsymbol{\alpha}|\boldsymbol{c}^{(s)},\boldsymbol{\beta}^{(s)},\boldsymbol{\Sigma}^{(s)},\tilde{\boldsymbol{y}}^{*(s-1)},\tilde{\boldsymbol{y}}) \end{aligned}$$

where $p(\boldsymbol{\pi}|\cdot)$, $p(\boldsymbol{\beta}|\cdot)$, $p(\boldsymbol{\Sigma}|\cdot)$ and $q(\boldsymbol{\alpha}|\cdot)$ correspond to the Gibbs steps of drawing $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$ and $\boldsymbol{\alpha}$, respectively. Often Metropolis-Hastings steps are used in the sampling of parameters such as $\boldsymbol{\alpha}$. In such a case, following Kaufmann and Frühwirth [36], $q(\boldsymbol{\alpha}|\cdot)$ now instead corresponds to the proposal density for drawing $\boldsymbol{\alpha}$ in a Metropolis-Hastings step. Assuming that we draw $\boldsymbol{c}$, $\boldsymbol{\pi}$, $\boldsymbol{\beta}$, $\boldsymbol{\Sigma}$, $\boldsymbol{\alpha}$ and $\tilde{\boldsymbol{y}}^*$ sequentially in the Gibbs algorithm, the values $\boldsymbol{c}^{(s)}$, $\boldsymbol{\beta}^{(s)}$ and $\boldsymbol{\Sigma}^{(s)}$ in 1.44 correspond to the s^{th} draw and $\boldsymbol{\Sigma}^{(s-1)}$ and $\tilde{\boldsymbol{y}}^{*(s-1)}$ the $(s-1)^{th}$ draw of the Gibbs algorithm. Frühwirth-Schnatter [21] demonstrates that the random permutation step forces the Gibbs algorithm to visit each of the $G!$ labeling subspaces. As a result, the importance density in (1.44) is a sampling average of densities with underlying parameters visiting all the $G!$ labeling subspaces. Without the random permutation step, the importance density in (1.44) may depend on parameters sticking in only one of the $G!$ labeling subspaces.

1.4.3 Instrument Imperfection

Procedures such as 2SLS and IV are commonly labeled as purely "classical" or "frequentist" and regarded as objective, free of the influences of priors. Despite this pervasive view, one can argue that such methods are, in fact, quite subjective and that beliefs surrounding the validity of such methods are rather personal.

To see this, consider a version of 1.1 and 1.2 where both $\boldsymbol{w}$ and $\boldsymbol{z}$ appear in the outcome and treatment equations:

$$\begin{aligned} y_i &= \beta_{y,0} + x_i\beta_{y,x} + \boldsymbol{w}_i\boldsymbol{\beta}_{y,w} + \boldsymbol{z}_i\boldsymbol{\beta}_{y,z} + \epsilon_i \qquad (1.45)\\ x_i &= \beta_{x,0} + \boldsymbol{w}_i\boldsymbol{\beta}_{x,w} + \boldsymbol{z}_i\boldsymbol{\beta}_{x,z} + u_i. \qquad (1.46) \end{aligned}$$

As argued in our introductory section, the model parameters are not fully identified without some form of additional structure. The approach behind instrumental variables is to determine a z or set of $\boldsymbol{z}$ for which $\boldsymbol{\beta}_{y,z} = 0$, and to achieve identification as a consequence of that restriction. In practice, we distinguish $\boldsymbol{w}$ from $\boldsymbol{z}$ via this belief.

What the IV-minded researcher does in practice, then, is to think carefully about the problem at hand and find a set of instruments which seem plausibly excludable. In papers of this type, an inordinate amount of time is spent in trying to convince the audience that it is reasonable to share his / her belief regarding this exclusion. But a *belief* is exactly what is being expressed here - an opinion of the investigator that an appropriate prior for the problem at hand is the dogmatic one in which $\boldsymbol{\beta}_{y,z} = 0$.

When thinking carefully about virtually all pursuits of this type based upon observational data, one can often come up with stories which undermine the perfect excludability of $\boldsymbol{z}$ from 1.45. For example, even quarter of birth, often used in past work to assess the causal impact of education on earnings, has recently been questioned as a legitimate instrument given its demonstrated correlation with maternal characteristics (Hungerman and Buckles [34]). This leads researchers to instead consider a system like 1.45 and 1.46, and to see what we might learn about the causal effect $\boldsymbol{\beta}_{y,x}$ when $\boldsymbol{\beta}_{y,z}$ is allowed to be "small" but non-zero. Recent work by Kraay [40] and Conley, Hansen and Rossi [18] discuss approaches based on this idea.

When allowing for instrument imperfection (that is, failing to impose $\beta_{y,z} = 0$), it is important to keep in mind the model becomes only partially identified. Given this partial identification, Chan and Tobias [5] note that, when even moderately weak priors on $\boldsymbol{\beta}_{y,z}$ are employed, standard Gibbs algorithms for this model can mix extremely poorly - in some cases millions or even billions of simulations may be required in order to obtain a desired level of numerical accuracy. To circumvent this problem, they outline a semi-analytic approach to the calculation of marginal posterior moments that even offers an improvement over the gold standard of iid sampling from the posterior.

Although generalized treatment-response models like 1.45 and 1.46 have substantial appeal, given the reality that instruments chosen in practice are not likely to be perfectly excludable, the resulting partial identification implies that the priors employed can have considerable influence. In fact, asymptotically, the conditional posterior for the non-identified structural parameters - like the causal effect $\beta_{y,x}$ - will reduce to the conditional prior (e.g., Moon and Schorfheide [51]), clearly revealing that the influence of the prior will not vanish as we acquire more data. As a result, it is incumbent upon the researcher to think very carefully about the priors selected, and perhaps to offer a menu of posterior results under a variety of different priors (one of which could be the common dogmatic choice $\boldsymbol{\beta}_{y,z} = 0$), in order to comprehensively document posterior sensitivity to different and reasonably maintained prior beliefs.

1.5 Illustrative Application

To illustrate how Bayesian methods can be used to estimate a treatment-response model, and how different models can be compared within that setting, we consider a version of the application conducted by Li, Mumford and Tobias [46].

This application considers analysis of data provided by an online payday loan lender, in operation across 38 different U.S. states. The variables to be modeled consist of the amount borrowed (typically around $300, with payment in full due within two weeks), as well as whether or not the borrower ultimately defaults on the loan. Default is a common phenomenon in this industry, and happens in approximately 30% of the loans in our data. We consider a triangular system in which the agent chooses an amount to borrow (perhaps to cover an emergency or unanticipated expense), and subsequently, the amount borrowed may have a causal effect on the decision to default.

States regulate a number of the terms of such loans, including the maximum interest rate that can be charged (*Rate*, and in practice, our lender generally follows the strategy of charging a rate equal to the state maximum rate), the maximum amount that can be borrowed ($StateMAX$), penalties that can be assessed on borrowers defaulting on a $300 loan (*StatePenalty*) as well as how long the lender (or collection agency) has to collect on the debt (*StatueLimit*). We use this variation in state policy to aid in identifying the parameters of our treatment-response model. Other variables used consist of the term of the loan (denoted $Term$, typically 14 days, as most borrowers in the sample are paid bi-weekly), and the monthly rent paid by the individual (*Rent*).

The specification we consider is given below:

$$\begin{aligned}
LogAmt_i^* &= \beta_{A,0c_i} + \beta_{A,1} Rate_i + \beta_{A,2} Term_i + \\
&\quad \beta_{A,3} LogRent_i + \beta_{A,4} LogStateMAX_i + \epsilon_i && (1.47)\\
LogAmt_i &= \min\{LogAmt_i^*, LogStateMAX_i\} && (1.48)\\
Default_i^* &= \beta_{D,0c_i} + \beta_{D,1} LogAmt_i + \beta_{D,2} Rate_i + \beta_{D,3} Term_i + \\
&\quad \beta_{D,4} LogRent_i + \beta_{D,5} LogStatePenalty_i + && (1.49)\\
&\quad \beta_{D,6} StatuteLimit_i + v_i \\
Default_i &= I(Default_i^* > 0) && (1.50)\\
\begin{pmatrix} \epsilon_i \\ v_i \end{pmatrix} \Bigg| \cdot &\overset{ind}{\sim} \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_{c_i} = \begin{pmatrix} \sigma_{\epsilon\epsilon c_i} & \sigma_{\epsilon v c_i} \\ \sigma_{v \epsilon c_i} & \sigma_{vv c_i} = 1 \end{pmatrix} \right] && (1.51)\\
\Pr(c_i = g) &= \pi_g, \text{ for } g = 1, 2, \cdots, G \text{ and } \sum_{g=1}^{G} \pi_g = 1. && (1.52)
\end{aligned}$$

Equation 1.47 can be interpreted as the amount that the agent would like to borrow, which is modeled as a function of the state-level interest rate, loan term, rent paid by the individual as well as the state-established maximum amount that can be borrowed. The desired amount is then linked to the observed (log) loan size through the restriction in 1.48 - actual amounts cannot exceed the state-established maximum, but otherwise are assumed to equal the desired loan size. (Provided the borrower meets a residency condition, a monthly income requirement, furnishes checking account information, and is not found to be present in a database of known defaulters, the requested loan amount is essentially approved by the lender).

Equation 1.49 can be interpreted as the net desire to default on the loan. The default decision is presumed to depend on loan size, the interest rate charged, the monthly rent paid by the individual and state-level penalties and collection terms on defaulting borrowers.

As discussed in our section on non-normality, we consider a finite mixture model, but allow only the intercepts and covariance matrices to vary across mixture components in order to avoid parameter proliferation. The notation above allows for G different mixture components; in practice we estimate specifications with up to 4 such components. Finally, note that the specification above represents a special case of the model described in 1.37 and 1.38, and as such, the Gibbs algorithm for that model can be readily applied here as well.

To determine the number of components to be included in the mixture modeling, we utilize the bridge sampling method explained in the previous section. In Table 1.1, we calculate the log marginal likelihood $\ln[p(\tilde{\boldsymbol{y}}|G\text{-component})]$ of a G-component mixture model, for $G = 1, 2, 3, 4$, and the Bayes factor $\frac{p(G\text{-component}|\tilde{\boldsymbol{y}})}{p(H\text{-component}|\tilde{\boldsymbol{y}})}$ of a G-component model as opposed to an H-component model, for $H = 1, 2, 3, 4$. The Bayes factors strongly suggest that the 2-component model is most favored by our payday loan data.

[Table 1 about here.]

Given this strong preference for a two-component specification, we focus our attention in the remainder of this discussion on parameter estimates obtained from this model. Posterior summary statistics for all model parameters are provided in Table 1.2.

Specifically, for each parameter, we report the posterior mean $\mathrm{E}(\theta_j|\tilde{\boldsymbol{y}})$, the posterior standard deviation $\mathrm{Std}(\theta_j|\tilde{\boldsymbol{y}})$, the posterior probability that the coefficient is positive $\Pr(\theta_j > 0|\tilde{\boldsymbol{y}})$, the numerical standard error (NSE) associated with the posterior mean and the marginal effect (ME) of the covariate associated with the parameter. These ME's are calculated for every observation in the sample and then averaged across those observations. For variables measured in dollars, the ME's represent changes resulting from a \$100 increase in the level of the given variable.

[Table 2 about here.]

These marginal impacts reveal interesting, and mostly expected results. For example, an increase in monthly rent of \$100 raises the expected loan amount by a modest \$1.39, while a similarly-sized \$100 increase in state borrowing limit leads to a much larger expected loan amount increase of \$11.10. The interest rate charged on the loan is found to have little effect on loan size, as the marginal posterior distribution of $\beta_{A,Rate}$ is nearly centered at zero with $\Pr(\beta_{A,Rate} > 0|\tilde{\boldsymbol{y}}) \approx .49$. This result is arguably consistent with the borrowers in our data seeking loans for emergency purposes, as that aspect of the loan seems to have little effect on loan size.

We also find that large loans are more likely to end in default, as a $100 increase in the amount borrowed raises the default probability by approximately 3 percentage points. Loans with higher interest rates are also associated with increased incidences of default, as a 0.1 increase in the interest rate increases the default probability by 4.84 percentage points. A $100 increase in monthly rent paid reduces the default probability by 1.04 percentage points, which is sensible given that individuals paying more in rent should be more likely to have the resources to pay off the loan when due. Finally, borrowers are sensitive to penalties associated with default; increasing the state-level penalty on default on a $300 loan by $100 decreases the default probability by 4.51 percentage points.

To focus on the "causal" effect of the amount borrowed on default probability, in Figure 1.1, we plot this marginal effect over the range of amount borrowed.

[Figure 1 about here.]

While the solid line represents the point estimates of the marginal effects, the dashed lines correspond to the posterior means plus and minus two times the posterior standard deviations. Note that the standard deviations are easy to calculate given the post-convergence simulations from the Gibbs sampler: for every point in the loan amount space, we obtain a collection of marginal effect values, one for each post-convergence simulation. This collection represents a sample from the induced marginal effect posterior distribution, from which standard deviations are easily calculated.

1.6 Conclusion

This chapter has provided a brief overview of parametric Bayesian methods for treatment response modeling. We first considered the fully linear triangular model and discussed MCMC implementation in that context. We then extended this to allow for discrete possibilities among both the treatment and response, illustrating the value of data augmentation and the connection between posterior simulators for linear models and those for nonlinear models conditioned on the latent data. We reviewed standard procedures for extending models beyond the usual Gaussian case and discussed a procedure for model comparison and selection. Finally, we illustrated how Bayesian calculations are performed and interpreted, using data from an online payday loan lender.

Applications of these methods abound in the literature, including, for example, Li [43], Li and Poirier [44], Munkin and Trivedi [52], Poirier and Tobias [55], Li, Poirier and Tobias[47], Deb, Munkin and Trivedi [19], [20], Hoogerheide, Kaashoek and van Dijk [33], Chib and Jacobi [13], [14], [15], Conley, Hansen and Rossi [18], Kline and Tobias [37], Hoogerheide, Block and Thurik [32] and Block, Hoogerheide and Thurik [2], among many others. Furthermore, endogeneity and causality in Bayesian economic modeling is discussed in many recent Bayesian textbooks, including Koop [38], Lancaster [42], Geweke [26], Rossi, Allenby and McCulloch [56], Koop, Poirier and Tobias [41], and Greenberg [29].

REFERENCES

1. Albert, J.H. and S. Chib (1993). "Bayesian Analysis of Binary and Polychotomous Response Data." *Journal of the American Statistical Association*, 88(422), 669-679.

2. Block, J.H., L.F. Hoogerheide and A.R. Thurik (2012). "Are Education and Entrepreneurial Income Endogenous? A Bayesian Analysis." *Entrepreneurship Research Journal*, 2(3).

3. Carlin, B.P. and N.G. Polson (1991). "Inference for Nonconjugate Bayesian Models Using the Gibbs Sampler." *The Canadian Journal of Statistics*, 19, 399-405.

4. Chan, J. C-C and I. Jeliazkov (2009). "MCMC Estimation of Restricted Covariance Matrices." *Journal of Computational and Graphical Statistics* 18(2), 457-480.

5. Chan, J. and J.L. Tobias (2013). "Priors and Posterior Computation in Linear Endogenous Variables Models with Imperfect Instruments." working paper.

6. Chib, S. (1992). "Bayes Inference in the Tobit Censored Regression Model." *Journal of Econometrics*, 51, 79-99.

7. Chib, S. (1995). "Marginal likelihood from the Gibbs output." *Journal of the American Statistical Association*, 90, 1313-1321.

8. Chib, S. (2007). "Analysis of Treatment Response Data Without the Joint Distribution of Potential Outcomes." *Journal of Econometrics*, 140, 401-412.

9. Chib, S. and E. Greenberg (2007). “Semiparametric Modeling and Estimation of Instrumental Variable Models.” *Journal of Computational and Graphical Statistics*, 16, 86-114.

10. Chib, S., E. Greenberg and I. Jeliazkov (2009). “Estimation of Semiparametric Models in the Presence of Endogeneity and Sample Selection.” *Journal of Computational and Graphical Statistics*, 18, 321-348.

11. Chib, S. and B. Hamilton (1999). “Bayesian Analysis of Cross-Section and Clustered Data Treatment Models.” *Journal of Econometrics*, 97, 25-50.

12. Chib, S. and B. Hamilton (2002). “Semiparametric Bayes Analysis of Longitudinal Data Treatment Models.” *Journal of Econometrics*, 110, 67-89.

13. Chib, S. and L. Jacobi (2007). “Modeling and Calculating the Effect of Treatment at Baseline from Panel Outcomes.” *Journal of Econometrics*, 140, 781-801.

14. Chib, S. and L. Jacobi (2008a). “Causal Effects from Panel Data in Randomized Experiments with Partial Compliance.” in S. Chib, W. Griffiths, G. Koop and D. Terrell (eds.), *Advances in Econometrics, Volume 23*, Bingley: Jai Press, 183-215.

15. Chib, S. and Jacobi, L. (2008b). “Analysis of Treatment Response Data from Eligibility Designs.” *Journal of Econometrics*, 144, 465-78.

16. Chib, S. and I. Jeliazkov (2001). “Marginal Likelihood from the Metropolis-Hastings Output.” *Journal of the American Statistical Association*, 96, 270-281.

17. Conley, T.G., C.B. Hansen, R.E. McCulloch and P.E. Rossi (2008). “A Semi-Parametric Bayesian Approach to the Instrumental Variables Problem.” *Journal of Econometrics*, 144(1), 276-305.

18. Conley, T.G., C.B. Hansen and P.E. Rossi (2012). “Plausibly Exogenous.” *Review of Economics and Statistics*, 94(1), 260-272.

19. Deb, P., M. Munkin and P.K. Trivedi (2006a). “Private Insurance, Selection, and the Health Care Use: A Bayesian Analysis of a Roy-Type Model.” *Journal of Business and Economic Statistics*, 24, 403-415.

20. Deb, P., M. Munkin and P.K. Trivedi (2006b). “Bayesian Analysis of the Two-Part Model with Endogeneity: Application to Health Care Expenditure.” *Journal of Applied Econometrics*, 21, 1081-1099.

21. Frühwirth-Schnatter, S. (2001). “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models.” *Journal of the American Statistical Association*, 96, 194-209.

22. Frühwirth-Schnatter, S. (2004). “Estimating Marginal Likelihoods for Mixture and Markov Switching Models Using Bridge Sampling Techniques.” *Econometrics Journal*, 7, 143-167.

23. Gelfand, A.E. and D.K. Dey (1994). “Bayesian Model Choice: Asymptotics and Exact Calculations.” *Journal of Royal Statistical Society B*, 56, 501-514.

24. Geweke, J. (1993). “Bayesian Treatment of the Independent Student-t Linear Model.” *Journal of Applied Econometrics*, 8, S19-S40.

25. Geweke, J. (1999). “Using Simulation Methods for Bayesian Econometric Models: Inference, Development, and Communication.” *Econometric Reviews*, 18, 1-73.

26. Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics.* John Wiley & Sons.

27. Geweke, J. (2007). "Interpretation and Inference in Mixture Models: Simple MCMC Works." *Computational Statistics and Data Analysis*, 51, 3529-3550.

28. 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.* Elsevier.

29. Greenberg, E. (2013). *Introduction to Bayesian Econometrics*, 2nd edition. Cambridge: Cambridge University Press.

30. Griffin, J., F. Quintana and M.F.J. Steel (2010). "Flexible and Nonparametric Modelling", in J. Geweke, G. Koop and H. van Dijk (eds.), *Handbook of Bayesian Econometrics.* Oxford: Oxford University Press.

31. Heckman, J. and B. Honoré (1990). "The Empirical Content of the Roy Model." *Econometrica*, 58(5), 1121-1149.

32. Hoogerheide, L.F., J.H. Block and A.R. Thurik (2012). "Family Background Variables as Instruments for Education in Income Regressions: A Bayesian Analysis." *Economics of Education Review*, 31(5), 515-523.

33. Hoogerheide, L.F., J.F. Kaashoek and H.K. van Dijk (2007). "On the Shape of Posterior Densities and Credible Sets in Instrumental Variable Regression Models with Reduced Rank: An Application of Flexible Sampling Methods Using Neural Networks." *Journal of Econometrics*, 139, 154-180.

34. Hungerman, D. and C. Buckles (2013). "Season of Birth and Later Outcomes: Old Questions, New Answers." *Review of Economics and Statistics*, forthcoming.

35. Jeliazkov, I., J. Graves and M. Kutzbach (2008). "Fitting and Comparison of Models for Multivariate Ordinal Outcomes." in S. Chib, W. Griffiths, G. Koop and D. Terrell (eds.), *Advances in Econometrics, Volume 23*, Emerald Group Publishing Limited, 115-156.

36. Kaufmann, S. and S. Frühwirth-Schnatter (2002). "Bayesian Analysis of Switching ARCH Models." *Journal of Time Series Analysis*, 23, 425-458.

37. Kline, B. and J.L. Tobias (2008). "The Wages of BMI: Bayesian Analysis of a Skewed Treatment Response Model with Nonparametric Endogeneity." *Journal of Applied Econometrics*, 23, 767-793.

38. Koop, G. (2003). *Bayesian Econometrics.* John Wiley & Sons.

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

40. Kraay, A. (2012). "Instrumental Variables Regressions with Uncertain Exclusion Restrictions: A Bayesian Approach." *Journal of Applied Econometrics*, 27, 108-128.

41. Koop, G., D.J. Poirier and J.L. Tobias (2007). *Bayesian Econometric Methods.* Cambridge: Cambridge University Press.

42. Lancaster, A. (2004). *An Introduction to Modern Bayesian Econometrics.* Blackwell Publishing.

43. Li, K. (1998). "Bayesian Inference in a Simultaneous Equations Model with Limited Dependent Variables." *Journal of Econometrics*, 85, 387-400.

44. Li, K. and D.J. Poirier (2003). “An Econometric Model of Birth Inputs and Outputs for Native Americans.” *Journal of Econometrics*, 113(2), 337-361.

45. Li, M. (2006). “High School Completion and Future Youth Unemployment: New Evidence from High School and Beyond.” *Journal of Applied Econometrics*, 21(1) 23-53.

46. Li, M., K. Mumford and J.L. Tobias (2012). “A Bayesian Analysis of Payday Loans and Their Regulation.” *Journal of Econometrics*, 171(2), 205-216.

47. 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, 203-25.

48. Lindley, D.V. and A.F.M. Smith (1972). “Bayes Estimates for the Linear Model.” *Journal of the Royal Statistical Society, Series B*, 34(1), 1-41.

49. McCulloch, R.E., N.G. Polson and P.E. Rossi (2000). “A Bayesian Analysis of the Multinomial Probit Model with Fully Identified Parameters.” *Journal of Econometrics*, 99, 173-193.

50. Meng, X.-L. and W.H. Wong (1996). “Simulating Ratios of Normalizing Constants via a Simple Identity: A Theoretical Exploration.” *Statistica Sinica*, 6, 831-860.

51. Moon, H. and F. Schorfheide (2012). “Bayesian and Frequentist Inference in Partially Identified Models.” *Econometrica*, 80(2), 755-782.

52. Munkin, M.K. and P.K. Trivedi (2003). “Bayesian Analysis of Self-Selection Model with Multiple Outcomes Using Simulation-Based Estimation: An Application to the Demand for Healthcare.” *Journal of Econometrics*, 114, 197-220.

53. Neal, R. (2000). “Markov Chain Sampling Methods for Dirichlet Process Mixture Models.” *Journal of Computational and Graphical Statistics*, 9(2), 249-265.

54. Nobile, A. (2000). “Comment: Bayesian Multinomial Probit Models with a Normalization Constraint.” *Journal of Econometrics*, 99, 335-345.

55. 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, 258-268.

56. Rossi, P.E., G.M. Allenby and R. McCulloch (2005). *Bayesian Statistics and Marketing*. John Wiley & Sons Ltd.

57. Roy, A.D. (1951). “Some Thoughts on the Distribution of Earnings.” *Oxford Economic Papers*, 3(2), 135-146.

58. Sims, C. (2007). “Thinking about Instrumental Variables.” available online at http://sims.princeton.edu/yftp/IV/.

59. Tanner, M. and W.H. Wong (1987). “The Calculation of Posterior Distributions by Data Augmentation.” *Journal of the American Statistical Association*, 82, 528-49.

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

Figure 1.1: Marginal effects of loan amount on default probability

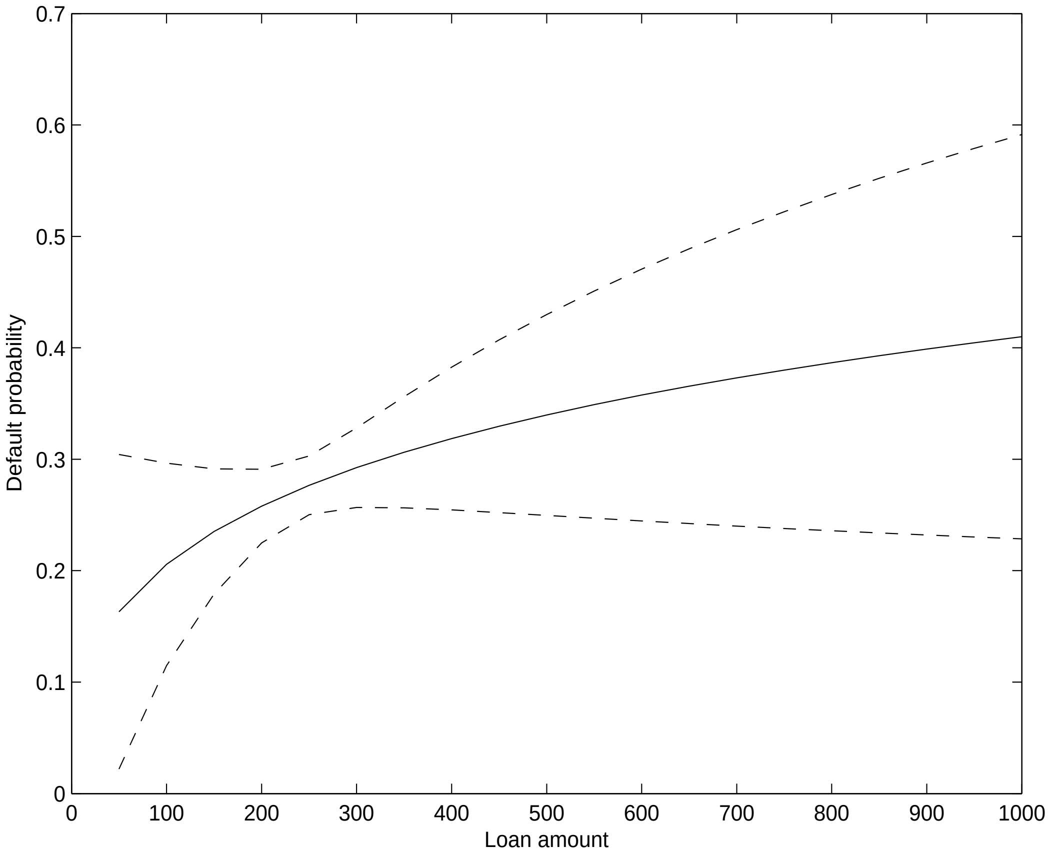

Table 1.1: Log marginal likelihoods and Bayes factors of G-component mixture models, for $G = 1, 2, 3, 4$

G	$\ln[p(\tilde{\boldsymbol{y}}\|G)]$	$\frac{p(\tilde{\boldsymbol{y}}\|D)}{p(H=1\|\tilde{\boldsymbol{y}})}$	$\frac{p(G\|\tilde{\boldsymbol{y}})}{p(H=2\|\tilde{\boldsymbol{y}})}$	$\frac{p(G\|\tilde{\boldsymbol{y}})}{p(H=3\|\tilde{\boldsymbol{y}})}$	$\frac{p(G\|\tilde{\boldsymbol{y}})}{p(H=4\|\tilde{\boldsymbol{y}})}$
1	-3338.8	1	8.7407×10^{-11}	1.3620×10^{-8}	5.4111×10^{-6}
2	-3315.6	1.1441×10^{10}	1	155.82	61907
3	-3320.7	7.3422×10^{7}	0.0064176	1	397.29
4	-3326.7	1.8481×10^{5}	1.6153×10^{-5}	0.002517	1

Table 1.2: Posterior summary statistics from the two-component model

Parameter (θ_j)	E($\theta_j \vert \tilde{\boldsymbol{y}}$)	Std($\theta_j \vert \tilde{\boldsymbol{y}}$)	Pr($\theta_j > 0 \vert \tilde{\boldsymbol{y}}$)	NSE	ME
$\beta_{A,Rate}$	-0.00506	0.298	0.492	0.00161	-0.0727
$\beta_{A,Term}$	0.00212	0.00223	0.83	1.23e-005	0.543
$\beta_{A,LogRent}$	0.0483	0.0283	0.957	0.000157	1.39
$\beta_{A,LogStateMAX}$	0.0712	0.0297	0.992	0.000162	11.1
$\beta_{D,LogAmt}$	0.382	0.266	0.929	0.00706	0.0296
$\beta_{D,Rate}$	1.91	0.905	0.982	0.00758	0.0484
$\beta_{D,Term}$	0.0189	0.00768	0.994	8.01e-005	0.00453
$\beta_{D,LogRent}$	-0.384	0.101	5e-005	0.00102	-0.0104
$\beta_{D,LogStatePenalty}$	-0.194	0.0493	0	0.000875	-0.0451
$\beta_{D,StatuteLimit}$	-0.00928	0.0124	0.226	9.04e-005	-0.00221
π_1	0.363	0.043	1	0.00156	
$\beta_{A,01}$	4.42	0.271	1	0.00147	
$\beta_{D,01}$	0.996	1.36	0.773	0.0273	
$\sigma_{\epsilon\epsilon 1}$	0.103	0.0103	1	0.000244	
$\rho_{\epsilon v1} = \sigma_{\epsilon v1}/\sqrt{\sigma_{\epsilon\epsilon 1}}$	0.0312	0.0865	0.644	0.00136	
π_2	0.637	0.043	1	0.00156	
$\beta_{A,02}$	5.02	0.276	1	0.00164	
$\beta_{D,02}$	-0.62	1.57	0.352	0.036	
$\sigma_{\epsilon\epsilon 2}$	0.285	0.0207	1	0.000518	
$\rho_{\epsilon v2} = \sigma_{\epsilon v2}/\sqrt{\sigma_{\epsilon\epsilon 2}}$	0.0934	0.116	0.79	0.00254	