Bayesian Econometric Methods

Joshua Chan
Purdue University
chan196@purdue.edu

Justin L. Tobias
Purdue University
jltobias@purdue.edu

February 2020

keywords: Bayes, Markov Chain Monte Carlo, Gibbs sampling, posterior simulation, probit model, treatment-response model, endogeneity, nonparametric estimation, Dirichlet process

---

Abstract

This chapter provides an introduction to the use of Bayesian methods in labor economics and related disciplines. Since the observed growth in Bayesian methods over the last two decades has largely been driven by computational advances, this passage focuses primarily on illustrating how such computations are performed in a selection of models that are relevant to applied work in this area. The chapter begins by discussing posterior simulation via Markov Chain Monte Carlo methods in the context of a binary choice model, which also contains an application involving Body Mass Index (BMI) and analysis of the likelihood of being overweight. Next, computation (i.e., posterior simulation) is discussed in a specification commonly encountered in applied microeconomics: a treatment-response model or, more specifically, a linear model with an endogenous right-hand side variable. The chapter closes with comments on the likely future of this literature, including a discussion and application of nonparametric Bayesian methods via the Dirichlet process.

---

1 Introduction

When asking many economists about the differences between Bayesian and frequentist methods for estimation and inference, most will probably cite the use of prior information by the former and a lack of adopting such information by the latter. While the authors of this chapter agree that Bayesian methods provide a formal role for the adoption of prior information and a coherent framework for updating beliefs upon the arrival of data, they disagree that, in most cases, frequentist methods are completely prior-free. In order to write any empirical paper, one must decide, among other things, what set of covariates to potentially include in an analysis, what model specification to adopt, whether or not to be concerned about problems of heteroscedasticity, endogeneity, or error of measurement, etc. If you were to isolate ten different researchers in individual rooms, each equipped with the same data set and each researcher posed with the same question, would you anticipate all would arrive at the same answer, even if only classical estimation methods were employed? It seems obvious, at least to this set of authors, that you would get a menu of different point estimates and answers, as the models and methods employed by each researcher would be the end result of a series of personal choices - choices often not documented or fully scrutinized when results of the study are presented or published. Curiously, people often object when priors are placed over parameters of a given model - priors that are typically non-informative relative to the data and yield the same asymptotic properties as frequentist estimates - but don't question foundational modeling choices that are arguably far more important, as they cannot be revised by the data.

At some level, the key distinction between Bayesian and frequentist inference is deeper than the prior (or lack of prior), even if you were to argue that priors are likely at play in both paradigms. Bayesian inference conditions on the observed data while frequentist inference involves averaging over data sets that may potentially have been observed, but were not. The prior is a means to the end of obtaining results that condition on the observed data.

In practice, the growing popularity of Bayesian methods in economics appears to have little to do with a warming of the empirical community to the ideology of Bayes, but rather, owes to the development of a variety of simulation-based tools that enable estimation and inference in models that might otherwise be intractable. Foremost among these tools are so-called Markov Chain Monte Carlo (MCMC) methods, which include the Gibbs sampler (see, e.g. Casella and George (1992)) and Metropolis-Hastings algorithms (see, e.g., Chib and Greenberg (1995)). Indeed, in some areas

of economics, such as branches of empirical macro economics and finance, Bayesian methods have grown to become the dominant framework in the profession.

Labor economics and related disciplines have also witnessed considerable growth in the application of Bayesian methods. For example, a number of studies in health economics, including Li and Poirier (2003a), (2003b), Munkin and Trivedi (2003), Geweke, Gowrisankaran and Town (2003) Deb, Munkin and Trivedi (2006), (2006a) Bretteville-Jensen and Jacobi (2009) Hu, Munkin and Trivedi (2015) and Jacobi and Sovinsky (2016), among others, have employed a Bayesian methodology. Other studies, including Li, Poirier and Tobias (2003), Koop and Tobias (2004), Kline and Tobias (2008), Li and Tobias (2011), Li, Mumford and Tobias (2012), Hoogerheide, Block and Thurik (2012), Block, Hoogerheide and Thurik (2012) Früwirth-Schnatter et al (2012), (2016), (2018), and Jacobi, Wagner and Früwirth-Schnatter (2016) have also explored labor-related topics. Researchers wanting to know more about Bayesian methods beyond the limited introduction in this chapter will find a wealth of information in popular texts on the subject, including Koop (2003), Lancaster (2004), Geweke (2005) and Greenberg (2008). Chapter 14 of Koop, Poirier and Tobias (2007) and Chan, Koop, Poirier and Tobias (2019) in particular, contains a detailed treatment of Bayesian estimation of popular models in applied microeconometric work.

The purpose of this chapter is to review some of the basic machinery for Bayesian posterior inference and to illustrate how those techniques work in models relevant for labor / microeconometric research and labor-related applications. The following section shows how MCMC can be used to fit a standard binary choice model. In section 3 that coverage is extended to review Bayesian estimation of a model that is very widely used in labor applications: a standard treatment - response model, i.e., a model with a right-hand side variable that is considered endogenous. Section 4 presents some very recent Bayesian developments, including a discussion and application of nonparametric Bayes based upon a Dirichlet process. Such models are particularly appealing in empirical labor work, as they enable the researcher to flexibly model response (parameter) heterogeneity and flexibly represent distributions without relying on rigid functional forms. The chapter concludes with a summary in section 5.

2 Bayesian Methods for Binary Choice: The Probit

Not surprisingly, Bayesian analysis of the probit model begins with an application of Bayes' theorem. This theorem states that the posterior distribution of some unobserved vector of model parameters,

say $\boldsymbol{\beta}$, is proportional to the product of the prior for those parameters, denoted as $p(\boldsymbol{\beta})$, and the likelihood, denoted as $p(\boldsymbol{y}|\boldsymbol{\beta})$. Formally,

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta})p(\boldsymbol{y}|\boldsymbol{\beta}). \tag{1}$$

For those unfamiliar to Bayes, the delivery of results up to proportionality may seem strange and possibly uncomfortable. The right-hand side of (1) is known and can be calculated: The researcher specifies a functional form for the prior $p(\boldsymbol{\beta})$, such as a normal distribution, and the likelihood is assumed available. If the product of prior and likelihood in the right-hand side can be graphed, this gives the shape of the posterior distribution, and the (unknown) normalizing constant is simply the value that scales the posterior properly so that it integrates to one. This normalizing constant can be difficult to calculate in problems of even moderate complexity, and that value often turns out to be key to Bayesian approaches to model selection and comparison. In principle, however, this normalizing constant can be determined; this determination turns out to be particularly easy when the right-hand side is recognized as the kernel of a known distribution.

In the case of a probit model considered in this section, the likelihood function is well-known:

$$p(\boldsymbol{y}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i}(1-\Phi(\boldsymbol{x}_i\boldsymbol{\beta}))^{1-y_i}\right], \tag{2}$$

thus yielding the joint posterior distribution:

$$p(\boldsymbol{\beta}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} \left[\Phi(\boldsymbol{x}_i\boldsymbol{\beta})^{y_i}(1-\Phi(\boldsymbol{x}_i\boldsymbol{\beta}))^{1-y_i}\right]. \tag{3}$$

In linear models, such as the familiar regression model, Gaussian likelihoods are known to combine nicely with Gaussian priors to yield Gaussian posteriors. In such cases the prior is said to be conjugate (or conditionally conjugate). The likelihood and prior in (3), however, do not combine nicely due to nonlinearity of the binary choice model. As such, the right-hand side of (3) is difficult to evaluate directly: it is hard, for example, to directly calculate posterior means, posterior standard deviations and other quantities of interest for the components of $\boldsymbol{\beta}$. The binary choice model is therefore very useful as an intoductory example of modern Bayesian estimation and inference, as it leads us to a discussion of powerful computational tools - useful in all kinds of models relevant for labor economics and related fields - that facilitate estimation and inference when direct calculation is either difficult or simply intractable.

The probit model also can be represented in terms of a latent variable z_i:

$$z_i = \boldsymbol{x}_i\boldsymbol{\beta} + \epsilon_i, \qquad \boldsymbol{\epsilon}|\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_n), \tag{4}$$

$$y_i = I(z_i > 0). \tag{5}$$

In the above, $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian (or normal) distribution with mean μ and variance σ^2. We previously mentioned that Gaussian linear models with Gaussian priors combine naturally; the representation of the probit in (4) is seen as a linear model *conditioned on* $\boldsymbol{z}$. This equivalent representation of the model and the computational conveniences associated with conditioning on $\boldsymbol{z}$ lead us to think about working with an *augmented posterior distribution* that treats the latent data $\boldsymbol{z}$ in a similar manner to unknown coefficient vector $\boldsymbol{\beta}$:

$$\begin{aligned} p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) &\propto p(\boldsymbol{y}, \boldsymbol{\beta}, \boldsymbol{z}) \\ &= p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta})p(\boldsymbol{z}, \boldsymbol{\beta}) \\ &= p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta})p(\boldsymbol{z}|\boldsymbol{\beta})p(\boldsymbol{\beta}). \end{aligned}$$

Thus, the augmented posterior distribution (which also includes $\boldsymbol{z}$) is proportional to the term that links the observed and latent data, $p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta})$, the "likelihood function" as if $\boldsymbol{z}$ were observed, $p(\boldsymbol{z}|\boldsymbol{\beta})$, and the prior for $\boldsymbol{\beta}$. The first of these terms is easy to determine, upon reflection: given $\boldsymbol{z}$, the system in (4) reveals that the values of $\boldsymbol{y}$ are known with certainty and $\boldsymbol{\beta}$ is superfluous: If z_i is positive, then $y_i = 1$ and if z_i is less than or equal to zero, then $y_i = 0$. Given the assumed independence across observations, the conditional density of $\boldsymbol{y}$ given $\boldsymbol{z}$ can be written as:

$$p(\boldsymbol{y}|\boldsymbol{z}, \boldsymbol{\beta}) = p(\boldsymbol{y}|\boldsymbol{z}) = \prod_{i=1}^{n} \big(I(y_i = 0)I(z_i \leq 0) + I(y_i = 1)I(z_i > 0)\big).$$

The term $p(\boldsymbol{z}|\boldsymbol{\beta})$ is implied by (4):

$$p(\boldsymbol{z}|\boldsymbol{\beta}) = \prod_{i=1}^{n} \phi(z_i; \boldsymbol{x}_i\boldsymbol{\beta}, 1),$$

where $\phi(x; \mu, \sigma^2)$ denotes the normal pdf for x with mean μ and variance σ^2. Putting all of this together, the augmented joint posterior follows:

$$p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{\beta}) \prod_{i=1}^{n} \left([I(y_i = 0)I(z_i \leq 0) + I(y_i = 1)I(z_i > 0)]\, \phi(z_i; \boldsymbol{x}_i\boldsymbol{\beta}, 1)\right). \tag{6}$$

The reader may pause at this point and ask why the representation in (6) does anything other than introduce unnecessary complications. If the posterior $p(\boldsymbol{\beta}|\boldsymbol{y})$ isn't easily characterized, then surely the augmented posterior in (6) must be even more unwieldy!

While this is true, there remains a subtle advantage to working with this larger, augmented posterior distribution. First, estimation will proceed by generating a series of draws from (6) rather than to try and directly calculate its moments or other posterior features. These draws can then be used to approximate whatever posterior statistic is desired: for example, the posterior mean of an element of $\boldsymbol{\beta}$ can be approximated as the average of the samples of that element that are drawn from the posterior. Second, it is recognized that *conditioned on* $\boldsymbol{z}$ the model is linear and one can easily generate samples from the conditional posterior distribution $p(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{z})$. This is essentially one-half of what is required of the Gibbs sampler, a simulation scheme that iteratively cycles through complete posterior conditional distributions of the model in order to generate samples from the joint posterior distribution (see, e.g., Casella and George (1992)). The remaining distribution necessary in this endeavor is $p(\boldsymbol{z}|\boldsymbol{y}, \boldsymbol{\beta})$. In what follows, both of these distributions are described in detail.

From (6),

$$p(\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) \propto p(\boldsymbol{\beta})\phi(\boldsymbol{z}; \boldsymbol{X}\boldsymbol{\beta}, \boldsymbol{I}_n).$$

This, as mentioned before, is simply a Gaussian linear model (since it conditions on $\boldsymbol{z}$) together with a prior for $\boldsymbol{\beta}$. Suppose the following prior is employed:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_\beta, \boldsymbol{V}_\beta).$$

This combination of prior and likelihood is well-known to produce a normal conditional posterior distribution (see, e.g., Lindley and Smith (1972) and Chan et al (2019), exercise 12.9), via completion of the square in $\boldsymbol{\beta}$:

$$\boldsymbol{\beta}|\boldsymbol{z}, \boldsymbol{y} \sim \mathcal{N}(\boldsymbol{D}_\beta \boldsymbol{d}_\beta, \boldsymbol{D}_\beta), \tag{7}$$

where

$$\boldsymbol{D}_\beta = \left(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{V}_\beta^{-1}\right)^{-1}, \quad \boldsymbol{d}_\beta = \boldsymbol{X}'\boldsymbol{z} + \boldsymbol{V}_\beta^{-1}\boldsymbol{\mu}_\beta.$$

For the remaining conditional posterior distribution, (6) again implies:

$$p(\boldsymbol{z}|\boldsymbol{\beta}, \boldsymbol{y}) \propto p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y}) \propto \prod_{i=1}^{n} \left(\left[I(y_i = 0) I(z_i \leq 0) + I(y_i = 1) I(z_i > 0) \right] \phi(z_i; \boldsymbol{x}_i\boldsymbol{\beta}, 1) \right).$$

Inspection of this expression reveals that each z_i can be sampled independently (since the joint conditional distribution is separable as a product of z_i terms) and, specifically,

$$p(z_i|\boldsymbol{\beta}, \boldsymbol{y}) \propto \left[I(y_i = 0) I(z_i \leq 0) + I(y_i = 1) I(z_i > 0) \right] \phi(z_i; \boldsymbol{x}_i\boldsymbol{\beta}, 1).$$

That is, conditionally, z_i has a normal distribution, but its support is truncated by the observed value of y_i. When $y_i = 1$, $z_i > 0$ and when $y_i = 0$, $z_i \leq 0$. Formally,

$$z_i|\boldsymbol{y}, \boldsymbol{\beta} \sim \begin{cases} \mathcal{TN}_{(0,\infty)}(\boldsymbol{x}_i\boldsymbol{\beta}, 1) & \text{if } y_i = 1 \\ \mathcal{TN}_{(-\infty,0]}(\boldsymbol{x}_i\boldsymbol{\beta}, 1) & \text{if } y_i = 0 \end{cases}, \quad i = 1, 2, \ldots, n, \tag{8}$$

where, notationally, $x \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2)$ denotes that x is a normally distributed random variable with (untruncated) mean μ and (untruncated) variance σ^2 which is then truncated to the interval (a, b). This truncated density retains the shape of the normal density over (a, b), is zero outside this interval, and is simply scaled up to integrate to one.

While one can generate draws from the truncated normal above by repeatedly drawing from a $\mathcal{N}(\boldsymbol{x}_i\boldsymbol{\beta}, 1)$ distribution and simply waiting for a draw that falls in the desired orthant, this process is quite inefficient. Samples from (8) can, however, be generated via the method of inversion (see, e.g., Chan et al (2019), exercises 11.4 and 11.5) among other possibilities. Specifically, let

$$u \sim U(0, 1)$$

be a draw from the uniform distribution on the unit interval. We can then form the variable w, where

$$w = \mu + \sigma\Phi^{-1}\left(\Phi\left(\frac{a-\mu}{\sigma}\right) + u\left[\Phi\left(\frac{b-\mu}{\sigma}\right) - \Phi\left(\frac{a-\mu}{\sigma}\right)\right]\right), \tag{9}$$

and simple derivations show that $w \sim \mathcal{TN}_{(a,b)}(\mu, \sigma^2)$.

Note that posterior simulation in the probit therefore involves only two steps, and each of these only involves sampling from a standard distribution (i.e., normal or truncated normal). A Gibbs algorithm for fitting the probit model proceeds as follows: Start with an initial value of the $\boldsymbol{\beta}$ vector. Given this initial $\boldsymbol{\beta}$ value, calculate $\boldsymbol{X\beta}$ and use this index to sample z_i, for $i = 1, 2, \ldots, n$ from (8), thus fully simulating the vector $\boldsymbol{z}$. Next, sample a new $\boldsymbol{\beta}$ (use the $\boldsymbol{z}$ just drawn to calculate an updated vector $\boldsymbol{d_\beta}$) from (7). The process repeats and converges to produce a set of (correlated) draws from the joint posterior $p(\boldsymbol{\beta}, \boldsymbol{z}|\boldsymbol{y})$. These post-convergence draws can then be used to calculate point estimates, standard deviations and other quantities of interest. An application of these methods is given in the following section.

2.1 Probit Model: Application

In this illustrative application, a data set from the British Cohort Study, as used by Kline and Tobias (2008), is analyzed. The primary question of interest surrounds the identification of characteristics

that are related to respondent Body Mass Index (BMI), and specifically, in predicting whether an individual is overweight, defined as a BMI in excess of 25. Code used in this analysis is available upon request, although the data source is restricted-use and thus cannot be similarly shared.

The analyzed sample consists of a set of male individuals with measured BMI (as opposed to self-reports of BMI or height and weight). Given BMI, the binary overweight indicator is constructed. Potential covariates include an intercept, BMI of the respondent's mother ($MotherBMI$) and father ($FatherBMI$), a marriage indicator ($Married$), an indicator denoting completion of a college degree ($Degree$) and an indicator denoting if the respondent reports to exercise regularly ($RegExercise$). The model is fit using Gibbs sampling, as described in the previous section, using the prior $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, 100\boldsymbol{I}_6)$. This corresponds to a prior that is quite flat (the marginal prior standard deviations all equal 10) and thus is weakly informative relative to the data. The sampler is run for 26,000 iterations and the first 1,000 of those are discarded as the burn-in period. The final 25,000 simulations are used to calculate parameter posterior means, standard deviations and the quantities reported in Table 1 below.

Table 1: Probit Analysis of BMI Data

	Coefficients			Marginal Effect	
Variable	$E(\cdot\|\boldsymbol{y})$	Std$(\cdot\|\boldsymbol{y})$	Pr$(\cdot > 0\|\boldsymbol{y})$	$E(\cdot\|\boldsymbol{y})$	Std$(\cdot\|\boldsymbol{y})$
Constant	-2.95	.260	0	—	—
FatherBMI	.069	.009	1.00	.028	.004
MotherBMI	.050	.007	1.00	.020	.003
Married	.240	.049	1.00	.095	.019
Degree	-.189	.061	.001	-.075	.024
RegExercise	.048	.064	.771	.019	.026

The first three columns of the table present coefficient posterior means, posterior standard deviations and posterior probabilities of being positive. Posterior means can be interpreted as point estimates of the model parameters, and commonly are reported in this way, although it's worth noting that different loss structures can and do give rise to point estimates other than the posterior mean. These point estimates are quantitatively very similar to frequentist maximum likelihood estimates in this case, which is to be expected with a moderate sample size and reasonably diffuse prior. Note that the third column of Table 1, the posterior probability that the coefficient is positive, is easily calculated from the given samples from the joint posterior distribution:

$$\widehat{\Pr(\beta_j > 0|\boldsymbol{y})} = \frac{1}{M}\sum_{m=1}^{M} I(\beta_{j,m} > 0),$$

where $\beta_{j,m}$ represents the m^{th} posterior simulation of the parameter β_j, $I(\cdot)$ denotes the standard indicator function and, in this case, $M = 25,000$. Note that this statistic answers a very useful question: what is the probability that the covariate of interest has a positive effect on the likelihood of being overweight? This quantity is easily explained and easily calculated from the posterior simulations and is often the misinterpretation of the classical $p-$value. The latter should not be interpreted as the probability that a hypothesis is false (or true), nor is such a question even well-posed in the frequentist paradigm. The $p-$value addresses the sampling question of how likely it would be to observe a value of a statistic that is at least as large as the one observed in the sample at hand, assuming the null is true. Such a question places a central role on data that could have been observed, but were not. The Bayesian approach, by contrast, conditions on the observed data and directly provides a quantity that, arguably, is more useful for practitioners and easily interpreted.

In terms of the results of the application, all coefficients, perhaps with the exception of $RegExercise$, seem to operate in the expected direction. Parental BMI associates positively with the likelihood of the respondent / child being overweight. Married respondents are more likely to be overweight, while those with a college degree are less likely to be overweight. As the third column indicates, results point to a high degree of confidence in the direction of these effects, as the posterior probabilities reported there are either one (i.e., all posterior simulations of that coefficient were positive) or near zero. The effect of regular exercise, however, is not precisely estimated. While one might expect the coefficient to be negative, it could also be the case that being overweight might lead someone to seek regular exercise.

In terms of marginal effects, a one-unit increase in father's BMI is associated with a 2.8 percent increase that the child is overweight. Marriage is associated with a 9.5 percent increase in the likelihood of being overweight, while those with a college degree are approximately 7.5 percent less likely to be overweight. Again, all of these calculations are easily performed given samples from the joint posterior. For example, the marginal effect of a continuous covariate is given as $\phi(\overline{x}\boldsymbol{\beta})\beta_j$, which can be calculated for each $\boldsymbol{\beta}$ drawn from the posterior. Taking an average of these quantities gives the fourth column of the table; the standard deviation provides the fifth.

Note as well that out-of-sample predictive quantities can be easily calculated. For example, suppose it is of interest to predict the likelihood of being overweight for a male who is married, does not exercise regularly, does not have a college degree and whose parents have BMIs of 30 (the clinical threshold for obesity). Stacking all of this information into a vector x_f, and letting y_f be the

associated unobserved overweight indicator, the following predictive probability is of interest:

$$\begin{aligned}\Pr(y_f = 1|x_f, \boldsymbol{y}) &= \int \Pr(y_f = 1|x_f, \boldsymbol{\beta}, \boldsymbol{y}) p(\boldsymbol{\beta}|\boldsymbol{y}) d\boldsymbol{\beta} \\ &= E_{\boldsymbol{\beta}|\boldsymbol{y}} \left[\Pr(y_f = 1|x_f, \boldsymbol{\beta}, \boldsymbol{y})\right],\end{aligned}$$

leading to

$$\widehat{\Pr(y_f = 1|x_f, \boldsymbol{y})} = \frac{1}{M} \sum_{m=1}^{M} \Phi(x_f \boldsymbol{\beta_m}).$$

Based upon the simulations obtained from the joint posterior, such a male has an 81 percent chance of being overweight (and the posterior standard deviation associated with $\Phi(x_f\boldsymbol{\beta})$ equals .024).

3 Endogeneity in Linear Models

The problem of endogeneity plays a central role in many, if not most, applications in labor economics and related disciplines. These applications often share a common structure: The causal effect of a key variable, say x, on some outcome y is sought, yet one recognizes that x is likely to be endogenous - factors unobserved by the econometrician are likely to be simultaneously related to both x and y (conditioned on x and other controls) . While most treatments of endogeneity in these literatures are classical in nature, centered upon or employing IV, 2SLS or other approaches for estimation, studies such as Drèze (1976), Geweke (1996), Kleibergen and Zivot (2003), Hoogerheide, Kleibergen and van Dijk (2007) and Conley et al (2008) mark important Bayesian advances to this literature. The importance of this issue is also suggested by the rather prominent and detailed treatment it receives in many current Bayesian textbooks (e.g., Lancaster (2004) - Chapter 8; Rossi, Allenby and McCulloch (2006) - Chapter 7; Koop, Poirier and Tobias (2007) and Chan, Koop, Poirier and Tobias (2019) - Chapter 14) . Furthermore, numerous applications have been tackled from a Bayesian point of view, often highlighting the ease with which MCMC methods can be adapted to deal with endogeneity problems in many different kinds of models [e.g., Li (1998); Geweke, Gowrisankaran and Town (2003); Li and Poirier (2003a), (2003b); Munkin and Trivedi (2003); Deb, Munkin and Trivedi (2006a); Kline and Tobias (2008); , Chib et al. (2009), Kraay (2012), Wiesenfarth et al (2014) and Chan and Tobias (2015)].

We discuss below a Bayesian treatment of endogeneity within the context of a linear regression model, where one of the right-hand side variables is endogenous. While this is somewhat restrictive, it is not terribly so, as simple generalizations can accommodate higher dimension endogeneity

problems. Moreover, a recent study by Chernozhukov and Hansen (2008) suggests that this is the modal model entertained in the literature and thus serves as a natural starting point for this analysis.

Consider as a starting point the model:

$$y_i = \alpha_0 + \alpha_1 x_i + \boldsymbol{\alpha}_2 \boldsymbol{w}_i + \epsilon_i \tag{10}$$

$$x_i = \beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i + u_i, \tag{11}$$

where

$$\left[\begin{array}{c} \epsilon_i \\ u_i \end{array} \right] \Bigg| \boldsymbol{W}, \boldsymbol{Z} \overset{iid}{\sim} \mathcal{N}\left[\left(\begin{array}{c} 0 \\ 0 \end{array} \right), \left(\begin{array}{cc} \sigma_\epsilon^2 & \sigma_{\epsilon u} \\ \sigma_{\epsilon u} & \sigma_u^2 \end{array} \right) \right] \equiv \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}).$$

Interest typically focuses on the first equation and the primary object of interest is α_1, the so-called causal effect of x on y. Endogeneity here is synonymous with $\sigma_{\epsilon u} \neq 0$; unobserved factors simultaneously correlate with x and y, leading to a confounding problem when trying to extract the causal effect. The exogenous variables $\boldsymbol{w}_i$ are covariates that enter the y-outcome equation while $\boldsymbol{z}_i$ enter the reduced form equation for x. As shown below, there can be (and almost always is) overlap between these two sets of variables, yet identification will require the appearance of at least one column of $\boldsymbol{Z}$ that is not contained in $\boldsymbol{W}$. This identification issue is discussed in more detail below.

Letting $\boldsymbol{\theta}$ denote all the parameters of the model, the joint density of the errors can be decomposed into a conditional times a marginal:

$$p(\epsilon_i, u_i|\boldsymbol{\theta}) = p(\epsilon_i|u_i, \boldsymbol{\theta}) p(u_i|\boldsymbol{\theta}). \tag{12}$$

Noting that the Jacobian of the transformation from (ϵ_i, u_i) to (y_i, x_i) is unity, the joint density of (y_i, x_i) follows:

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi\left(y_i \left| \alpha_0 + \alpha_1 x_i + \boldsymbol{\alpha}_2 \boldsymbol{w}_i + \frac{\sigma_{\epsilon u}}{\sigma_u^2}(x_i - \beta_0 - \boldsymbol{\beta}_1 \boldsymbol{z}_i), \sigma_\epsilon^2(1 - \rho_{\epsilon u}^2) \right.\right) \\ &\quad \times \phi(x_i|\beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i, \sigma_u^2), \end{aligned} \tag{13}$$

where $\rho_{\epsilon u} \equiv \sigma_{\epsilon u}/[\sigma_\epsilon \sigma_u]$.

Note that (13) provides the likelihood function (or, at least, one observations' contribution to the likelihood function). It is instructive to pause and discuss identification in the context of this system of equations. To this end, first consider the case where the set of exogenous covariates are common

to both equations, i.e., $\boldsymbol{z}_i = \boldsymbol{w}_i$. In this case, (13) becomes:

$$\begin{aligned} p(y_i, x_i|\boldsymbol{\theta}) &= \phi\left(y_i \left| \left[\alpha_0 - \beta_0 \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] + \left[\alpha_1 + \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] x_i + \left[\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_1 \frac{\sigma_{\epsilon u}}{\sigma_u^2}\right] \boldsymbol{w}_i, \sigma_\epsilon^2(1-\rho_{\epsilon u}^2)\right.\right) \\ &\quad \times \phi(x_i|\beta_0 + \boldsymbol{\beta}_1 \boldsymbol{z}_i, \sigma_u^2). \end{aligned} \tag{14}$$

Some quick accounting, then, shows that the likelihood is a function of just 7 (blocks of) parameters:

$$\beta_0,\ \boldsymbol{\beta}_1,\ \sigma_u^2,\ \psi_0 = [\alpha_0 - \beta_0 \frac{\sigma_{\epsilon u}}{\sigma_u^2}],\ \psi_1 = [\alpha_1 + \frac{\sigma_{\epsilon u}}{\sigma_u^2}],\ \boldsymbol{\psi}_2 = [\boldsymbol{\alpha}_2 - \boldsymbol{\beta}_1 \frac{\sigma_{\epsilon u}}{\sigma_u^2}] \text{ and } \psi_3 = \sigma_\epsilon^2(1-\rho_{\epsilon u}^2), \tag{15}$$

whereas the model is comprised of 8 distinct "structural" parameters:

$$\alpha_0, \quad \alpha_1, \quad \boldsymbol{\alpha}_2, \quad \beta_0, \quad \boldsymbol{\beta}_1, \quad \sigma_u^2, \quad \sigma_\epsilon^2, \text{ and } \sigma_{\epsilon u}. \tag{16}$$

As a result, the quantities in (15) are identified by the likelihood yet the full set of structural parameters in (16) is not identifiable. The researcher is, essentially, one equation short when seeking to recover this information from the data. Importantly, note that the "causal effect" α_1 - the object that garners most attention in practice - is among the parameters that are not identifiable when the set of covariates appearing in $\boldsymbol{w}$ and $\boldsymbol{z}$ are the same.

While several assumptions regarding the model can be used to achieve identification in this setting, the most common one is to assume the presence of at least one element of $\boldsymbol{z}$ that is not contained in $\boldsymbol{w}$. That is, a careful understanding of the problem at hand leads to the determination of a set of variables in $\boldsymbol{z}_i$ that are not contained in $\boldsymbol{w}_i$ and can be exploited for purposes of identification and estimation. Indeed, (13) shows how such exclusion restrictions can be exploited for identification purposes: The parameter $\boldsymbol{\beta}_1$ is identifiable from the marginal (reduced form) density of x_i, and the coefficient on the elements of $\boldsymbol{z}$ *not contained in* $\boldsymbol{w}$ in the conditional density $y|x$ becomes $-[\sigma_{\epsilon u}/\sigma_u^2]\boldsymbol{\beta}_1$. Together, these two pieces of information enable identification of the ratio $\sigma_{\epsilon u}/\sigma_u^2$, which is attributable to the role of unobserved confounding. Once this ratio is known, the causal effect α_1 as well as the remaining parameters of the model clearly become identifiable, as is evident from (13). This simple argument illustrates the value of instruments as vehicles for identification, and also suggests potential difficulties in separating α_1 from $\sigma_{\epsilon u}/\sigma_u^2$ when the instruments are poor (weak).

3.0.1 Posterior Simulation in the Linear Endogenous Variable Model

Given the availability of a valid instrument, the variables of the model can be first stacked into vectors and matrices by writing:

$$\begin{bmatrix} y_i \\ x_i \end{bmatrix} = \begin{bmatrix} 1 & x_i & \boldsymbol{w}_i & 0 & \mathbf{0} \\ 0 & 0 & \mathbf{0} & 1 & \boldsymbol{z}_i \end{bmatrix} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \boldsymbol{\alpha}_2 \\ \beta_0 \\ \boldsymbol{\beta}_1 \end{bmatrix} + \begin{bmatrix} \epsilon_i \\ u_i \end{bmatrix} \tag{17}$$

or

$$\tilde{\boldsymbol{y}}_i = \tilde{\boldsymbol{X}}_i \boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}_i, \tag{18}$$

with $\tilde{\boldsymbol{y}}_i$, $\tilde{\boldsymbol{X}}_i$, $\boldsymbol{\beta}$ and $\tilde{\boldsymbol{\epsilon}}_i$ defined in the obvious ways. Furthermore, suppose priors of the following forms are employed:

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}) \tag{19}$$

$$\boldsymbol{\Sigma}^{-1} \sim W\left[(\kappa \boldsymbol{R})^{-1}, \kappa\right]. \tag{20}$$

The first of these is the familiar normal (or Gaussian) distribution, while the latter may be less familiar to some readers. The assumed prior for the inverse covariance matrix is a Wishart prior, with degrees of freedom parameter κ and scale matrix $\boldsymbol{R}$. The Wishart distribution can be thought of as a multivariate generalization of the chi-square distribution, and several routines exist for generating draws from the Wishart.

With this setup in hand, posterior simulation in the linear model with endogeneity follows in a straightforward way. In particular, a simple two-block Gibbs algorithm can be employed that iteratively samples from the following two conditional posterior distributions:

$$\boldsymbol{\beta} | \boldsymbol{\Sigma}, \boldsymbol{y}, \boldsymbol{x} \sim \mathcal{N}(\boldsymbol{D}_{\boldsymbol{\beta}} \boldsymbol{d}_{\boldsymbol{\beta}}, \boldsymbol{D}_{\boldsymbol{\beta}}), \tag{21}$$

where

$$\boldsymbol{D}_{\boldsymbol{\beta}} = \left(\boldsymbol{V}_{\boldsymbol{\beta}}^{-1} + \sum_{i=1}^{n} \tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{X}}_i \right)^{-1}, \quad \boldsymbol{d}_{\boldsymbol{\beta}} = \boldsymbol{V}_{\boldsymbol{\beta}}^{-1} \boldsymbol{\mu}_{\boldsymbol{\beta}} + \sum_{i=1}^{n} \left(\tilde{\boldsymbol{X}}_i{}' \boldsymbol{\Sigma}^{-1} \tilde{\boldsymbol{y}}_i \right) \tag{22}$$

and

$$\boldsymbol{\Sigma}^{-1} | \boldsymbol{\beta}, \boldsymbol{y}, \boldsymbol{x} \sim W \left(\left[\sum_{i=1}^{n} \tilde{\boldsymbol{\epsilon}}_i \tilde{\boldsymbol{\epsilon}}_i{}' + \kappa \boldsymbol{R} \right]^{-1}, n + \kappa \right). \tag{23}$$

A posterior simulator for this model proceeds by iteratively sampling from (21) and (23). As mentioned previously, both of these are easily obtained. Note that a sweep of the posterior simulator

thus provides a sample of all of the structural parameters ($\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$) as well as the variance and covariance parameters via $\boldsymbol{\Sigma}$. In classical approaches, discussion of the reduced-form equation for $\boldsymbol{x}$ is typically muted, if not non-existent, apart perhaps from summary F-statistics that speak to the strength of the instrument. Posterior simulations of $\sigma_{\epsilon u}$ and the associated correlation $\rho_{\epsilon u}$ (which is easily calculated at each iteration) can also be used to speak to the degree of the endogeneity problem in the application at hand; values of $\rho_{\epsilon u}$ close to zero signal that endogeneity is weak and single-equation analyses of y on x and $\boldsymbol{w}$ are likely to give similar results, while values closer to one in absolute value signal the importance of controlling for endogeneity.

The model discussed here can easily be extended to account for, among other possibilities: multiple endogenous right-hand side variables and outcomes / endogenous variables that are discrete-valued. For the latter, latent variable representations again prove to be computationally useful, as described in the previous section. The references listed at the outset of this section provide examples of such generalizations. However, the canonical model analyzed in this section still relies on the assumption of jointly normal error terms, an assumption that may be violated in practice and is unlikely to satisfy the critical eyes of referees. The following section discusses, in general terms, nonparametric Bayesian approaches to allow for flexible modeling of distribution functions.

4 An Introduction to Nonparametric Bayesian Modeling

In early applications of Bayesian econometric methods, particularly those prior to the computational revolution that began in the early 1990s, it was common to assume that error terms followed a particular parametric distribution. This practice may, in fact, remain common, at least for early stages of exploratory data analysis.

As discussed earlier in this chapter, assuming that errors are normally distributed can be an appealing assumption to make, as it often proves mathematically convenient for posterior analysis when coupled with the adoption of normal priors. Such rigid forms, however, are typically based on computational convenience rather than application-at-hand-appropriateness and don't fully allow the data to speak for themselves, but instead require the researcher to (partially) speak on its behalf. The general trend in applied econometric work in the profession is a movement toward robustness and remaining as agnostic as possible regarding modeling assumptions. Bayesian methods have followed suit, and it seems quite reasonable to believe that future applied work in this area will continue to be characterized by the adoption of flexible semiparametric and nonparametric

methods. What follows is a discussion and illustration of one approach to Bayesian nonparametric / semiparametric modeling, focusing on the Dirichlet Process.

The Dirichlet Process (DP) and associated mixture model for observational data, termed the Dirichlet Process Mixture Model (DPMM), offers an avenue for flexible modeling of distributions within the Bayesian paradigm. The DP prior can be thought of as a prior distribution over a space of distributions. Prior hyperparameters in this endeavor specify the base distribution - the expected distributional realization - and a concentration parameter that controls how tightly any distributional draws will track the base distribution. As the prior is updated by the data, beliefs are revised about the overall shape of the distribution. This, of course, differs in a fundamental way from simply learning about a finite vector of parameters within an assumed distributional family.

The charge of this chapter is not to review specific technical details of DP modeling, but below a brief introduction to the methods and an illustrative application are provided. The reader is referred to seminal papers by Ferguson (1973), Sethuraman (1994) and Escobar and West (1995) for further technical details. Applications of the methodology that follows in models relevant to labor and related fields include Hirano (2002), Conley et al (2008), Chib and Greenberg (2010), Wiesenfarth et al (2014), Hu, Munkin and Trivedi (2015), Chan et al (2017) and Kim and Wang (2019), among others.

To explain a DPMM approach to nonparametric distributional modeling in broad terms, consider the following hierarchical system of equations:

$$\begin{aligned} y_i|\boldsymbol{\theta}_i &\sim F(\boldsymbol{\theta}_i) \\ \boldsymbol{\theta}_i|\boldsymbol{G} &\sim \boldsymbol{G} \\ \boldsymbol{G}|\alpha, \boldsymbol{G}_0 &\sim DP(\alpha, \boldsymbol{G}_0). \end{aligned}$$

The first equation of the above system looks like a typical parametric model that is commonly used for data analysis, but is more general given that each observation contains its own parameter vector $\boldsymbol{\theta}_i$. One can think of this first equation, like the example to be provided below, as a regression equation where parameters of the conditional mean function vary across each observation. Density estimation can be interpreted as a special case of this, where the "regression" just includes an intercept and variance parameter, both of which vary across observations.

The second equation puts a prior on the $\boldsymbol{\theta}_i$ and states that they are *iid* draws from some unknown distribution $\boldsymbol{G}$. The final equation places a prior on this unknown distribution and states that it

follows a Dirichlet Process with base distribution $\boldsymbol{G}_0$ and concentration parameter α. At this stage the technical details of the DP mixture model are far from clear - including what the DP actually is - and perhaps all that the reader can take away is the spirit of what the nonparametric Bayesian model seeks to do: put a prior over a space of prior distributions, and allow the data to inform the shape of that distribution in a way that is not restricted to a particular parametric family.

An alternate representation of this model may provide some useful intuition. Sethuraman (1994) shows that the DP can be represented as an infinite mixture. In the case where F is Gaussian, the sampling model can be represented as follows:

$$y_i|\boldsymbol{\omega}, \boldsymbol{\theta} \sim \sum_{j=1}^{\infty} \omega_j \mathcal{N}(y_i; \boldsymbol{\theta_j}). \tag{24}$$

In the above, $\mathcal{N}(x; \boldsymbol{\delta})$ represents the normal distribution for x with parameters $\boldsymbol{\delta}$, and the ω_j are weights associated with the respective mixture components. These can be constructed via a "stick-breaking" process whereby remaining portions of a stick of unit length are sequentially broken off. Specifically, this equivalent representation of the model in (24) follows from Sethuraman's (1994) constructive definition of the DP:

$$\boldsymbol{G} = \sum_{j=1}^{\infty} \omega_j \delta_{\boldsymbol{\theta}_j}, \quad \boldsymbol{\theta}_j \sim \boldsymbol{G}_0, \quad \eta_j \sim \text{Beta}(1, \alpha), \quad \omega_j = \eta_j \prod_{l<j} (1 - \eta_l), \tag{25}$$

which shows that realizations from the DP are discrete with probability one, and the mass points / atoms θ_j are drawn from the base distribution $\boldsymbol{G}_0$ with weights ω_j constructed from the stick breaking procedure. In practice, the Dirichlet Process exhibits a clustering property whereby the $\boldsymbol{\theta}_i$ tend to concentrate on $M < n$ different values. Thus, in DP modeling, the data are used to determine an appropriate finite mixture without having to specify the number of needed mixture components. One can therefore think of the Dirichlet process as allowing for parameter (and distributional) heterogeneity, and the ability to learn about the nature of that heterogeneity without requiring specific parametric forms to model its generation.

4.1 Application of Nonparametric Bayesian Modeling: Returns to Schooling

To illustrate application of the DPM model and reveal its potential use in uncovering unobserved heterogeneity in models of interest in labor economics and related fields, a small illustrative application is introduced involving returns to education. The data set analyzed consists of a .25% random sample of observations from the 2016 American Community Survey. Specifically, outcomes

of 888 working-age males who are employed full-time are analyzed. The specification considered is given below:

$$y_i = \beta_{0i} + \beta_{1i} Educ_i + \sigma_i \epsilon_i, \quad \boldsymbol{\epsilon} | \boldsymbol{Educ} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n), \tag{26}$$

where y_i is the log wage received by individual i and $Educ_i$ refers to years of schooling completed. Each observation in (26), at least potentially, is associated with its own intercept, slope and variance parameter. Although this specification seems at first glance to be unwieldy, excessively parameter rich and simply not identifiable, structure is added by employing a Dirichlet process prior to $\boldsymbol{\theta}_i = [\beta_{0i} \quad \beta_{1i} \quad \sigma_i]$ to induce clustering and parsimony to the analysis, and to let the data speak to the degree of response heterogeneity. The standard homogenous return-to-schooling model would, of course, be produced if the analysis collapses to a single cluster, and the common β_1 (abstracting from obvious endogeneity concerns) would be interpreted as the treatment effect / return to education. Departures from this single component model suggest the presence of heterogeneity in the data.

The prior specification continues as follows:

$$\begin{aligned} \boldsymbol{\theta}_i | \boldsymbol{G} &\sim \boldsymbol{G} \\ \boldsymbol{G} &\sim DP(\alpha, \boldsymbol{G}_0). \end{aligned}$$

That is, the $\boldsymbol{\theta}_i$ are assumed to follow some unknown distribution $\boldsymbol{G}$ and a DP prior is placed over this unknown distribution. The distribution $\boldsymbol{G}_0$ represents the prior mean of $\boldsymbol{G}$, and it is selected as the familiar independent normal, inverse gamma distribution:

$$\boldsymbol{G}_0 = \mathcal{N}(\boldsymbol{\beta}; \boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{V}_{\boldsymbol{\beta}}) \times IG(\sigma^2; a, b)$$

with

$$\boldsymbol{\mu}_{\boldsymbol{\beta}} = [3 \ .1]', \quad \boldsymbol{V}_{\boldsymbol{\beta}} = \begin{bmatrix} (2^2) & 0 \\ 0 & (.01)^2 \end{bmatrix}, \quad a = 2.49, \quad b = 1.92.$$

The choices of a and b set the prior mean of σ^2 to .35, with an associated prior standard deviation of .5. For the hyperparameters of the β, prior, the intercept of the base measure is centered at 3 and the return to schooling parameter is centered at .1, with standard deviations equal to 2 and .1, respectively.

The parameter α tends to govern how many clusters or components will be present in the data. Large values of α tend to produce many mass points with small weights. In these cases, draws from $\boldsymbol{G}$ tend to resemble those of the base distribution $\boldsymbol{G_0}$ and the model will tend toward one with a large number of components. On the other hand, realizations of $\boldsymbol{G}$ under small values of

α will tend to produce just a few atoms with large weight assigned to them, and thus the model will tend to concentrate on a small number of clusters or components. In practice, α is treated as a parameter to be updated from the data rather than fixed *a priori*.

Details of the MCMC algorithm are not provided here, although it is noted that slice sampling is employed to generate samples from the DP model. Interested readers are referred to Kalli, Griffin and Walker (2011) for further details. The sampler is run for 50,000 iterations, and the first 10,000 of those are discarded as the burn-in period. Code and the data set employed in this analysis are available upon request.

To begin, results associated with the number of distinct clusters (or mixture components) that are supported by the data are provided. The number of active clusters is denoted as M, which is produced for each post-convergence simulation. The table below provides posterior frequencies associated with distinct values of M:

Table 2: Posterior Frequencies of M

M	$\Pr(M = \cdot \vert \boldsymbol{y})$
1	0
2	.57
3	.30
4	.09
5	.02

As the table shows, no posterior simulations are associated with the benchmark single-component Gaussian model, signaling a strong preference for some form of heterogeneity in outcomes across individuals. Furthermore, most of the posterior mass concentrates around the adoption of 2 or 3 mixture components, and models that are more parameter-rich do not receive considerable support from the data. Table 3 below provides more details associated with the regression and variance parameters in these two and three component cases.

Table 3: Posterior Summary Statistics
Associated with 2 and 3 Component Models

	$M = 2$		$M = 3$		
Intercept (β_0)	1.37	1.84	1.37	1.78	2.51
	(.171)	(.970)	(.169)	(.970)	(1.48)
Educ (β_1)	.122	.025	.122	.025	.074
	(.012)	(.055)	(.012)	(.056)	(.085)
σ^2	.295	.648	.294	.579	.386
	(.028)	(.362)	(.025)	(.337)	(.421)
ω	.919	.081	.926	.068	.006

Considering first those posterior simulations associated with the $M = 2$ case, it is seen that one component receives a majority of the weight (approximately 92% of the weight). In this high-weight component, the posterior mean estimate of the return to education is .122, consistent with a range of findings in the literature. The smaller-weight second component is associated with a significantly lower, though much noisier, return to schooling estimate, as well as a substantially higher conditional variance σ^2. Moving on to the $M = 3$ component case, a very similar pattern emerges. Specifically, one component receives a majority of the weight, and posterior results associated with that component are virtually identical to the high-weight component in the $M = 2$ case. The remaining 8 percent (or so) of the mass is then divided over two components, both of which are again associated with smaller returns to education and larger conditional variances.

These results point toward the overall need to control for heterogeneity, as $M = 1$ receives no support from the data. A precise description of that heterogeneity, however, remains difficult to fully characterize given that small-weight components in the $M > 1$ cases are associated with both higher variances and lower returns to education. To shed some additional light on this issue, the model is re-estimated, but this time β_0 and σ^2 are imposed to be common across observations and the DP prior is placed over the return to schooling parameter only. When doing so, results change substantially: the single-component ($M = 1$) specification now receives a majority of the weight (i.e., $\Pr(M = 1|\boldsymbol{y}) = .64$), and most of the remaining mass is spread over the $M = 2$ and $M = 3$ component cases ($\Pr(M = 2|\boldsymbol{y}) = .23$, $\Pr(M = 3|\boldsymbol{y}) = .08$). The two component case is again associated with a high-weight component ($E(\omega|\boldsymbol{y}) = .94$) associated with a mean return to schooling equal to .115 and a lower-weight component ($E(\omega|\boldsymbol{y}) = .06$) associated with a return equal to .094. However, differences between these components is not stark, as the marginal posterior distributions overlap considerably and the higher-weight component is associated with a larger return to education only 68 percent of the time. A similar pattern is found within the $M = 3$

case, as posterior means of β_1 (and associated component weights w) are .116 (.880), .113 (.109) and .078 (.011). Taken together, these results point toward a model with homogenous returns to schooling: the single-component specification receives the most support from the data and results allowing for additional heterogeneity often produce similar values of the slope parameter. The primary reason for preferring heterogeneity in the more general model of Table 1 appears to be its allowance for higher-variance outcomes, through heterogeneity in both β_0 and σ^2.

5 Summary

This chapter reviewed Bayesian approaches to the estimation of some models useful for research in labor economics and related disciplines. As much of the appeal of Bayesian methods in modern applications surrounds the computational tools they employ, this discussion began with an illustration of such tools in a simple binary choice setting. The chapter continued by showing how Markov Chain Monte Carlo computational methods (namely, the Gibbs sampler) can be employed to estimate a standard treatment-response model - a specification that underlies much work in this area. Finally, the topic of nonparametric Bayesian modeling via the Dirichlet Process was introduced, revealing how application of such methods can flexibly model unknown distributions and uncover underlying parameter heterogeneity.

References

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).

Bretteville-Jensen, A. L., and L. Jacobi (2009): "Climbing the Drug Staircase: A Bayesian Analysis of the Initiation of Hard Drug Use," *Journal of Applied Econometrics*, 23, 1157–1186.

Casella, G., and E. George (1992): "Explaining the Gibbs Sampler," *The American Statistician*, 46, 167–174.

Chan, J., D. J. Henderson, C. F. Parmeter, and J. L. Tobias (2017): "Nonparametric Estimation in Economics; Bayesian and Frequentist Approaches," *WIREs Computational Statistics*.

Chan, J., G. Koop, D. J. Poirier, and J. L. Tobias (2019): *Bayesian Econometric Methods*. Cambridge University Press, second edn.

Chan, J., and J. L. Tobias (2015): "Priors and Posterior Computation in Linear Endogenous Variables Models with Imperfect Instruments," *Journal of Applied Econometrics*, 30, 650–674.

Chernozhukov, V., and C. Hansen (2008): "The Reduced Form: A Simple Approach to Inference with Weak Instruments," *Economics Letters*, 100, 68–71.

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

——— (2010): "Additive Cubic Spline Regression with Dirichlet Process Mixture Errors," *Journal of Econometrics*, 156, 322–336.

Conley, T., C. Hansen, R. McCulloch, and P. Rossi (2008): "A Semi-Parametric Bayesian Approach to the Instrumental Variables Problem," *Journal of Econometrics*, 144, 276–305.

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

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

Dréze, J. (1976): "Bayesian Limited Information Analysis of the Simultaneous Equations Model," *Econometrica*, 44, 1045–1075.

Escobar, M. D., and M. West (1995): "Bayesian Density Estimation and Inference Using Mixtures," *Journal of the American Statistical Association*, 90, 577–588.

Ferguson, T. (1973): "A Bayesian Analysis of Some Nonparametric Problems," *Annals of Statistics*, 1, 209–230.

Frühwirth-Schnatter, S., C. Pamminger, A. Weber, and R. Winter-Ebmer (2012): "Labor Market Entry and Earnings Dynamics: Bayesian Inference Using Mixtures-of-Experts Markov Chain Clustering," *Journal of Applied Econometrics*, 27, 1116–1137.

——— (2016): "Mothers' Long-Run Career Patterns After First Birth," *Journal of the Royal Statistical Society, Series A*, 179, 707–725.

Frühwirth-Schnatter, S., S. Pittner, A. Weber, and R. Winter-Ebmer (2018): "Analysing Plant Closure Effects Using Time-Varying Mixture-of-Experts Markov Chain Clustering," *Annals of Applied Statistics*, 12, 1786–1830.

Geweke, J. (1996): "Bayesian Reduced Rank Regression in Econometrics," *Journal of Econometrics*, 75, 121–146.

——— (2005): *Contemporary Bayesian Econometrics and Statistics*. Wiley.

Geweke, J., G. Gowrisankaran, and R. J. Town (2003): "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71(4), 1215–1283.

Greenberg, E. (2008): *Introduction to Bayesian Econometrics*. Cambridge University Press.

Hirano, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781–799.

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.

Hoogerheide, L. F., F. Kleibergen, and H. H. van Dijk (2007): "Natural Conjugate Priors for the Instrumental Variables Regression Model Applied to the Angrist-Krueger Data," *Journal of Econometrics*, 138, 63–103.

Hu, X., M. Munkin, and P. K. Trivedi (2015): "Estimating Incentive and Selection Effects in the Medigap Insurance Market: An Application with Dirichlet Process Mixture Model," *Journal of Applied Econometrics*, 30, 1115–1143.

Jacobi, L., and M. Sovinsky (2016): "Marijuana on Main Street?' Esimtating Demand in Markets with Limited Access," *American Economic Review*, 106, 20092045.

Jacobi, L., H. Wagner, and S. Früwirth-Schnatter (2016): "Bayesian Treatment Effects Models with Variable Selection for Panel Outcomes with an Application to Earnings Effects of Maternity Leave," *Journal of Econometrics*, 193, 234–250.

Kalli, M., J. E. Griffin, and S. G. Walker (2011): "Slice Sampling Mixture Models," *Statistics and Computing*, 21, 93–105.

Kim, J., and L. Wang (2019): "Hidden Group Patterns in Democracy Developments: Bayesian Inference for Grouped Heterogeneity," *Journal of Applied Econometrics*, 34, 1016–1028.

Kliebergen, F., and E. Zivot (2003): "Bayesian and Classical Approaches to Instrumental Variable Regression," *Journal of Econometrics*, 114, 29–72.

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

Koop, G. (2003): *Bayesian Econometrics*. Wiley.

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

Koop, G., and J. L. Tobias (2004): "Learning About Heterogeneity in Returns to Schooling," *Journal of Applied Econometrics*, 19, 723–747.

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

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

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

Li, K., and D. J. Poirier (2003a): "The Roles of Birth Inputs and Outputs in Predicting Health, Behavior, and Test Scores in Early Childhood," *Statistics in Medicine*.

——— (2003b): "An Econometric Model of Birth Weight for Native Americans," *Journal of Econometrics*.

Li, M., K. Mumford, and J. L. Tobias (2012): "Bayesian Analysis of Payday Loans and Their Regulation," *Journal of Econometrics*, 171, 205–216.

Li, M., D. J. Poirier, and J. L. Tobias (2003): "Do Dropouts Suffer from Dropping Out? Estimation and Prediction of Outcome Gains in Generalized Selection Models," *Journal of Applied Econometrics*, 19, 203–225.

Li, M., and J. L. Tobias (2011): "Bayesian Inference in a Correlated Random Coefficients Model: Modeling Treatment Effect Heterogeneity and Heterogeneous Returns to Schooling," *Journal of Econometrics*, 162, 346–361.

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

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 Health Care," *Journal of Econometrics*, 114(2), 197–220.

Rossi, P., G. Allenby, and R. McCulloch (2006): *Bayesian Statistics and Marketing*. Wiley.

S. Chib, E. G., 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.

Sethuraman, J. (1994): "A Constructive Definition of Dirichlet Priors," *Statistica Sinica*, 4, 639–650.

Wiesenfarth, M., C. M. Hisgen, T. Kneib, and C. Cadarso-Suarez (2014): "Bayesian Nonparametric Instrumental Variables Regression Based on Penalized Splines and Dirichlet Process Mixtures," *Journal of Business and Economic Statistics*, 32, 468–482.