Confronting Prior Convictions: On Issues of Prior Sensitivity and Likelihood Robustness in Bayesian Analysis

Hedibert F. Lopes[1] and Justin L. Tobias[2]

[1]Booth School of Business, University of Chicago, Chicago, Illinois 60637; email: hlopes@chicagobooth.edu

[2]Economics Department, Purdue University, Lafayette, Indiana 47907; email: jltobias@purdue.edu

Annu. Rev. Econ. 2011. 3:107–31

The *Annual Review of Economics* is online at economics.annualreviews.org

This article's doi: 10.1146/annurev-economics-111809-125134

JEL codes: C01, C11, C52

1941-1383/11/0904-0107$20.00

Keywords

Bayesian methods, marginal likelihood, scale mixture of normals, Dirichlet process mixture, factor models, Markov chain Monte Carlo, Gibbs sampler, sequential Monte Carlo

Abstract

In this review we explore issues of the sensitivity of Bayes estimates to the prior and form of the likelihood. With respect to the prior, we argue that non-Bayesian analyses also incorporate prior information, illustrate that the Bayes posterior mean and the frequentist maximum likelihood estimator are often asymptotically equivalent, review a simple computational strategy for analyzing sensitivity to the prior in practice, and finally document the potentially important role of the prior in Bayesian model comparison. With respect to issues of likelihood robustness, we review a variety of computational strategies for significantly expanding the maintained sampling model, including the use of finite Gaussian mixture models and models based on Dirichlet process priors.

1. INTRODUCTION

As any Bayesian practitioner will readily confess, it is rather common—and indeed expected—that when presenting a Bayesian study one will face questions regarding the influence of the prior and, perhaps to a lesser extent, the sensitivity of results to the form of the likelihood function. Although such concerns are certainly valid, and any serious Bayesian endeavor should document aspects of sensitivity with respect to the choice of prior and functional form, it remains unquestionably true that many who ask such questions do so out of an inherent distrust of the Bayesian paradigm or the lessons learned from its application. To say that priors are specific to the Bayesian is to deny the existence of such preconceived opinions in the profession—a position of belief that the Bayesian can, and potentially has, arrived at the result she wants only after a sufficient amount of meddling with the model and prior.

In this article we seek to dispel some of these prior convictions and discuss some issues regarding the sensitivity of posterior results to the prior. In addition, we review procedures for enabling flexible representations of the likelihood.

With respect to the issue of prior sensitivity and the Bayesian's use of prior information, we begin by arguing that empirical work in economics is inevitably colored by the thoughts and opinions of the researcher, with prior beliefs insinuated at various stages of the modeling process. To say then that the Bayesian approach is unique or clearly differentiated from the classical approach to inference because of its reliance on prior information is misleading. It is differentiated from the frequentist approach, however, in that, conditioned on a maintained model or set of models, priors are incorporated in a transparent and formal manner and added to data information in a way that conforms to basic rules of probability theory. We also note that in smooth, regular finite-dimensional models, Bayes and frequentist estimators are asymptotically equivalent and illustrate this equivalence in a simple exponential family sampling model with a conjugate prior.

Although these responses to prior-sensitivity concerns can be interpreted as somewhat defensive (as the spirit of such replies is to turn the question of prior sensitivity back on those asking it), it is nonetheless true that the prior can have an effect on the posterior. We therefore describe a simple method for assessing prior sensitivity when Markov chain Monte Carlo (MCMC) output is available and also discuss the important role of the prior in Bayesian model comparison.

With respect to likelihood robustness, we also review several computational strategies that generalize the Gaussian sampling models commonly associated with Bayesian work. This includes a discussion of scale and finite mixtures of Gaussian distributions, as well as analyses based on Dirichlet process (DP) priors. Importantly, these methods enable a flexible specification of the sampling model or prior distribution for a set of coefficients while maintaining computational tractability.

The outline of this review is as follows. Section 2 discusses the role of the prior. We discuss how priors are used, either formally or informally when conducting empirical research in economics, and then move on to illustrate, within the context of a natural exponential family sampling model, the asymptotic equivalence of the posterior mean and the maximum likelihood estimator (MLE). When simulation-based methods are used to fit a model, we also review a method to address the sensitivity of posterior means to the specification of the prior. We close Section 2 by noting an important instance for which the prior can be quite influential—in the computation of marginal likelihoods. Section 3 then discusses empirical strategies for generalizing the maintained sampling model, focusing on

scale and finite Gaussian mixture models and nonparametric Bayes via DP priors. The review concludes with a summary in Section 4.

2. ROBUSTNESS AND SENSITIVITY TO THE PRIOR

As convincingly summarized by Lancaster (2004, p. 8), the research process in economics and statistics is not truly objective, as priors and personal opinions, often absent from the published form of the study although instrumental in its construction, play a prominent role. To see this, let us consider a researcher faced with a question of interest and in possession of a data set that enables him to shed light on that question. Our researcher begins his analysis by making a number of decisions on how to best process the data and by combing through a myriad of different potential models. He must inevitably deal with a number of different issues, ranging from the selection of key variables to include in the analysis while identifying those that certainly can be discarded, the determination of the proper treatment of incomplete or missing data, the identification of the subpopulation to study and the external validity of the results, to the employment of an appropriate sampling model that captures the key features of the problem at hand.

Undoubtedly, two different individuals, faced with the same data set and seeking an answer to the same question, would arrive at different final models and point estimates for the quantities of interest. Conditioned on the model employed, of course, the analysis is packaged with a sheen of objectivity, as there seems to be no formal role for a prior. This is not to say, however, that the analysis has been prior-free: Our two classical researchers will not arrive at exactly the same answer, although it is hoped that they will reach some sort of qualitative agreement, and therefore the results are robust.

The popularity of robustness exercises and their prevalence in published empirical studies go hand in hand with concerns regarding the influences of priors—in this case, priors that have been used to select a model. Such robustness checks, although valuable, are necessarily limited, addressing only the sensitivity to select aspects of the modeling process, and focusing on elaborations of the model that the researcher himself (or perhaps a referee) deems to be most important.

Although it seems to us a rather indisputable fact that priors inevitably play a role in the research process, it seems equally true that many become uncomfortable when a prior distribution over a set of parameters is formally introduced as an ingredient of the Bayesian's model. Such a reaction to priors, however, must also surface at earlier stages of the modeling process—in discarding a set of variables from consideration, for example, the researcher is acting as if she has a dogmatic prior for the parameters of a more general model that restricts these parameters to be zero. In this sense, priors used in the selection of a model or class of models are far more informative than those that the Bayesian employs over the parameters of a given model, as data information can revise at least the latter type of prior, whereas no amount of data information can serve to revise the former type of belief. On this rather perplexing state of affairs, Tukey (1978, p. 52) writes

> It is my impression that rather generally, not just in econometrics, it is considered decent to use judgement in choosing a functional form but indecent to use judgement in choosing a coefficient. If judgement about important things is quite all right, why should it not be used for less important ones as well?[1]

[1]This quotation was brought to our attention by Poirier (1995, p. 524).

In the following section we document a particular sense in which the prior regarding the parameters of the model indeed can be regarded as less important. Specifically, we illustrate that sampling-based asymptotic inference for the posterior mean is identical to standard classical inference for the MLE, and in this sense the prior washes out in sufficiently large samples. It is also true that, under certain conditions, the (suitably scaled) posterior distribution itself converges to the asymptotic distribution of the MLE. The implications of this result are that Bayesian posterior intervals generally will enjoy good frequentist coverage properties, and conversely, reported confidence intervals, although fundamentally different in interpretation, are often very close numerically to Bayesian posterior intervals. We illustrate aspects of this in the next section, focusing on a reasonably general exponential sampling model for a scalar random variable, together with a conjugate prior for a scalar parameter of interest.

2.1. A Scalar Exponential Family Example

Consider a set of scalar independent and identically distributed (i.i.d.) observations, denoted here by $y_1, y_2, \ldots, y_n$, from the natural exponential family of distributions:

$$p(y_i \mid \theta) = h(y_i) \exp[\theta y_i - g(\theta)], \quad i = 1, 2, \ldots, n, \tag{1}$$

where the functions h and g are known, and $\mathbf{y}_n$ will be used to denote the full vector of n observations.

The usual recipe for sampling-based inference involves forming the likelihood function, determining the MLE, and then characterizing a large-sample approximation to its distribution in repeated samples. Before proceeding with these details, however, we observe (provided one can interchange the order of integration and differentiation)

$$\int_Y p_\theta(y \mid \theta) dy = \int_Y p(y \mid \theta)[y - g_\theta(\theta)] = 0, \tag{2}$$

where $p_\theta(y \mid \theta) \equiv \partial p(y \mid \theta)/\partial\theta$, and $g_\theta(\theta)$ is defined similarly. Equation 2 immediately implies

$$E(y) = g_\theta(\theta_0), \tag{3}$$

where for simplicity we assume that we have correctly specified the distributional family, θ_0 represents the true parameter of the data-generation process, and expectations are taken with respect to the actual data-generation process, $p(y \mid \theta = \theta_0)$. By extension,

$$\mathrm{Var}(y) = g_{\theta\theta}(\theta_0), \tag{4}$$

where $g_{\theta\theta}(\cdot)$ is analogously defined as the second derivative.

It is straightforward to show that the MLE is determined by solving

$$g_\theta(\hat{\theta}_n) = \bar{y}_n,$$

where $\bar{y}_n \equiv n^{-1} \sum_{i=1}^n y_i$, and $\hat{\theta}_n$ represents the MLE. That is, $g_\theta(\theta_0)$, from Equation 3, can be interpreted as the population average of the random variable y, and our maximum likelihood estimate is derived from the sample analog of this moment condition. For notational simplicity, let $\hat{\psi}_n \equiv g_\theta(\hat{\theta}_n)$ and $\psi_0 \equiv g_\theta(\theta_0)$. Under suitable regularity conditions, we obtain the following asymptotic approximation to the sampling distribution of $\hat{\psi}_n$:

$$\sqrt{n}(\hat{\psi}_n - \psi_0) \xrightarrow{d} \mathcal{N}[0, g_{\theta\theta}(\theta_0)], \tag{5}$$

from which large-sample confidence intervals can be constructed.

2.2. Bayesian Inference in the Exponential Family Sampling Model

The Bayesian takes the exponential family likelihood above, adds to it a prior, and combines them via Bayes' theorem to obtain a posterior distribution for θ. In this case, a conjugate prior (that is, one that yields a posterior distribution of the same functional form) is given by

$$p(\theta \mid \underline{a}, \underline{b}) \propto \exp[\underline{a}\theta - \underline{b}g(\theta)], \tag{6}$$

where $\underline{a}$ and $\underline{b}$ are hyperparameters selected by the researcher. Using the same type of argument that was used to derive Equation 3, it is straightforward to show, provided the moment exists,

$$E(\psi) \equiv \underline{\mu}_{\psi} = \underline{a}\,\underline{b}^{-1}, \tag{7}$$

with $\psi \equiv g_{\theta}(\theta)$. Combining the likelihood derived from Equation 1 with the prior in Equation 6, we obtain via Bayes' theorem

$$p(\theta \mid y_n) \propto \exp[(n\bar{y}_n + \underline{a})\theta - (n + \underline{b})g(\theta)]. \tag{8}$$

Inspection of Equation 8 reveals that the prior in Equation 6 is indeed conjugate, as the posterior in Equation 8 is of the same functional form as the prior, with updated definitions of the parameters $\underline{a}$ and $\underline{b}$, say, $\bar{a} = n\bar{y}_n + \underline{a}$ and $\bar{b} = n + \underline{b}$. Using the result in Equation 7, we then obtain

$$E(\psi \mid \mathbf{y}_n) = (n\bar{y}_n + \underline{a})(n + \underline{b})^{-1}, \tag{9}$$

which we write equivalently as

$$E(\psi \mid \mathbf{y}_n) = \omega_n \hat{\psi}_n + (1 - \omega_n)\underline{\mu}_{\psi}, \tag{10}$$

a weighted average of the data mean (or MLE $\hat{\psi}_n = \bar{y}_n$) and the prior mean $\underline{\mu}_{\psi}$, with the weight ω_n defined as $n/(n + \underline{b})$. Equation 10 immediately reveals

$$\sqrt{n}[E(\psi \mid y) - \psi_0] = \omega_n[\sqrt{n}(\hat{\psi}_n - \psi_0)] + O(n^{-1/2}).$$

As $\omega_n \to 1$, we see that the Bayesian posterior mean and the frequentist MLE in this case are asymptotically equivalent. That is, the asymptotic sampling properties of the Bayesian posterior mean are identical to the sampling properties of the classical MLE.

In our view this result places the frequentist who questions the role of the prior in a potentially precarious position. To conduct inference, for which finite-sample results are seldom available, the frequentist typically relies on asymptotic approximations to the sampling distribution of the estimator. But under this large-sample metric, the sampling distribution of the Bayes rule is identical, suggesting that, according to his own recipe for inference, the prior does not matter. To question the role of the prior then from a sampling theory perspective is to introduce the importance of finite-sample results; otherwise concerns about the prior are misplaced and unwarranted, as the Bayes rule simply represents a different estimator with identical large-sample sampling properties.[2]

This above example can be regarded as illustrating an aspect of the Bernstein–von Mises theorem, which, in smooth finite-dimensional models, establishes the closeness of the

[2]Admittedly, the finite-sample performances of various classical procedures have been and continue to be an active area of research, focusing on the application of bootstrap, jacknife, and other methods for improving finite-sample inference.

Bayes estimate to the MLE and the posterior distribution of the parameter around the mean with the sampling distribution of the MLE around the true value of the parameter.[3] With respect to this last point, not specifically illustrated in our example above, we mean that the posterior distribution of the quantity $\sqrt{n}[\psi - E(\psi \mid \boldsymbol{y}_n)]$ often converges to the same asymptotic normal distribution as in Equation 5. In this sense, Bayesian posterior intervals in moderate to large samples generally will enjoy good frequentist coverage probabilities, whereas, numerically, the reported classical confidence interval should be close to the Bayesian posterior interval. It is in this general sense that one can argue that the prior is less important and that, although fundamentally different in interpretation, Bayesian posterior and frequentist confidence intervals are often numerically similar and may be approximately interchangeable.

2.3. When the Prior Matters

The foregoing discussion was not intended to leave the reader with the impression that checking for prior sensitivity is an unnecessary exercise. Instead, our intent was to document that all analyses are inevitably influenced by the beliefs of the researcher and that from a sampling perspective, Bayes and frequentist estimators are often asymptotically equivalent.

Whereas the prior will tend to have minimal impact on the posterior distribution in a fixed finite-dimensional model with a moderate to large sample size, the prior can have an impact on posterior results when observations are scarce or few relative to the number of parameters. Below we describe how concerns regarding prior sensitivity can be addressed without needing to refit the model for each new prior under consideration.[4]

2.4. Assessing the Sensitivity of Posterior Means to the Prior

It probably comes as no surprise to the reader that most modern Bayesian empirical work employs simulation methods for model fitting. A variety of different iterative and noniterative methods, some of which are discussed in the following section, can be applied to generate a set of draws from the joint posterior distribution. These draws then can be used to approximate parameter posterior means, posterior standard deviations, and posterior quantiles or to quantify out-of-sample predictive outcomes when such quantities cannot be obtained analytically.

In this section we take it as given that a posterior simulator has been applied to fit a baseline model. We denote this baseline model as $\mathcal{M}_0$, which is characterized by the likelihood $p(\boldsymbol{y} \mid \theta, \mathcal{M}_0)$ and prior $p(\theta \mid \mathcal{M}_0)$. Furthermore, a set of posterior simulations, denoted $\{\theta_0^{(m)}\}$, $m = 1, 2 \ldots, M$, is generated from the posterior distribution $p(\theta \mid \boldsymbol{y}, \mathcal{M}_0)$, and an estimate of the posterior mean $E(\theta \mid \boldsymbol{y}, \mathcal{M}_0)$ is calculated as the sample average of these simulations. We focus here on a scalar θ for simplicity, but note that the discussion below obviously generalizes to the case of vector-valued θ.

[3]This connection is not established when dealing with infinite-dimensional models or models with many parameters relative to the number of observations [see Freedman 1999 for discussion; for more on asymptotic posterior analysis, see, for example, Bernardo & Smith 2000, particularly section 5.3, and the references therein (many of which are enumerated on pp. 288–89)].

[4]We also recognize that there are cases in which the prior can be quite influential, even with very large sample sizes. Popular examples of such cases include models containing weak instruments and, of course, models that are only partially identified (for more on Bayesian analysis involving instrumental variables and related methods, see, for example, Geweke 1996; Kleibergen & Zivot 2003; Rossi et al. 2005, chapter 7; Sims 2007; Conley et al. 2008).

The researcher seeks to determine how sensitive the reported posterior mean (the typical Bayesian point estimate) is to changes in the prior. One way to answer this question, of course, is to re-estimate the model under the new prior, to obtain simulations from this new posterior distribution, and use these simulations to again calculate the posterior mean. Such an approach, however, is potentially unappealing, as considerable effort and computing time may be required to fit the model again, and the researcher may wish to do so for a wide variety of different priors.

An alternative approach, in the spirit of importance sampling (see, e.g., Kloek & van Dijk 1978, Geweke 1989), is simply to reweight the simulations from the initial baseline model to assess the impact of the prior change. To see why such an approach is valid, and how it is performed in practice, consider a different model, denoted $\mathcal{M}_1$, that contains the same parameter θ and likelihood function that characterize the baseline model $\mathcal{M}_0$ yet that introduces a different prior $p(\theta \mid \mathcal{M}_1)$. We begin by noting that the posterior mean under this new prior is obtained as

$$E(\theta \mid \boldsymbol{y}, \mathcal{M}_1) = \frac{\int_\Theta \theta p(\boldsymbol{y} \mid \theta, \mathcal{M}_1) p(\theta \mid \mathcal{M}_1) d\theta}{\int_\Theta p(\boldsymbol{y} \mid \theta, \mathcal{M}_1) p(\theta \mid \mathcal{M}_1) d\theta}, \tag{11}$$

where the denominator represents the normalizing constant of the posterior distribution.

Now suppose we were to use the baseline posterior $p(\theta \mid \boldsymbol{y}, \mathcal{M}_0)$ as an importance function to numerically approximate values of both the numerator and denominator integrals in Equation 11. The insight of importance sampling is to divide and multiply terms within the integrand by another density—from which draws are easily obtained—to enable the application of direct Monte Carlo integration. The choice of the baseline posterior as the importance function affords some considerable simplifications to this general exercise, as the likelihoods are unchanged in $\mathcal{M}_1$ and $\mathcal{M}_0$, and the normalizing constant of the baseline posterior cancels in the ratio of Equation 11. Taking these things into account, the desired posterior mean can be written as

$$E(\theta \mid \boldsymbol{y}, \mathcal{M}_1) = \frac{\int_\Theta \theta [p(\theta \mid \mathcal{M}_1)/p(\theta \mid \mathcal{M}_0)] p(\theta \mid \boldsymbol{y}, \mathcal{M}_0) d\theta}{\int_\Theta [p(\theta \mid \mathcal{M}_1)/p(\theta \mid \mathcal{M}_0)] p(\theta \mid \boldsymbol{y}, \mathcal{M}_0) d\theta}. \tag{12}$$

The advantage of Equation 12 is that the averaging within the integrals is done now with respect to the baseline posterior $p(\theta \mid \boldsymbol{y}, \mathcal{M}_0)$, for which a set of simulations is already available. As a result, one does not need to re-estimate the model to assess prior sensitivity, but instead can simply reweight the baseline posterior simulations in the appropriate way. Specifically, the forms of the integrals in Equation 12 suggest that a simulation-consistent estimate of the new posterior mean is

$$E(\theta \mid \widehat{\boldsymbol{y}, \mathcal{M}_1}) = \sum_{m=1}^{M} \theta_0^{(m)} \omega_m, \tag{13}$$

where the weights ω_m are defined as

$$\omega_m \equiv \frac{p\left[\theta_0^{(m)} \mid \mathcal{M}_1\right] / p\left[\theta_0^{(m)} \mid \mathcal{M}_0\right]}{\sum_{m=1}^{M} p\left[\theta_0^{(m)} \mid \mathcal{M}_1\right] / p\left[\theta_0^{(m)} \mid \mathcal{M}_0\right]}.$$

In this way, one can easily assess how the posterior mean changes with changes in the prior by simply reweighting the initial $\{\theta_0^{(m)}\}_{m=1}^{M}$. Simulations that are more likely to arise under the new prior are appropriately assigned more weight in the calculation. The above weights also suggest that, provided the Bayesian documents her prior and makes her postconvergence simulations publicly available, any reader of the study can easily calculate revised posterior means under his subjective beliefs or can use this general strategy to investigate issues of robustness to changes in the prior.

2.5. Marginal Likelihoods and the Prior

In this section we move beyond estimation and review the role of the prior in the exercise of Bayesian model selection and comparison. To fix ideas, suppose that there are two competing models (either nested or nonnested) denoted $\mathcal{M}_1$ and $\mathcal{M}_0$. We note

$$p(\mathcal{M}_i \mid \boldsymbol{y}) = \frac{p(\boldsymbol{y} \mid \mathcal{M}_i)p(\mathcal{M}_i)}{p(\boldsymbol{y})}, \quad i = 0, 1,$$

and thus

$$K_{10} \equiv \frac{p(\mathcal{M}_1 \mid \boldsymbol{y})}{p(\mathcal{M}_0 \mid \boldsymbol{y})} = \left(\frac{p(\boldsymbol{y} \mid \mathcal{M}_1)}{p(\boldsymbol{y} \mid \mathcal{M}_0)}\right)\left(\frac{p(\mathcal{M}_1)}{p(\mathcal{M}_0)}\right). \tag{14}$$

Equation 14 reveals that the posterior odds of $\mathcal{M}_1$ relative to $\mathcal{M}_0$ (the left-hand side of the equation) equal the Bayes factor (the first ratio on the right-hand side) multiplied by the ratio of prior odds. Calculating the Bayes factor in Equation 14 can be difficult, as it involves calculating the marginal density of the data under each of the respective models, i.e.,

$$p(\boldsymbol{y} \mid \mathcal{M}_i) = \int_{\Theta_i} p(\boldsymbol{y} \mid \theta_i, \mathcal{M}_i)p(\theta_i \mid \mathcal{M}_i)d\theta_i, \quad i = 0, 1. \tag{15}$$

Under equal prior odds, the Bayes factor of Equation 14 summarizes the degree of support of one model over another, based on the given data. As Equation 15 clearly reveals, the priors for the parameters of the models play an important role in this process. In what follows we illustrate a potentially surprising result connecting the priors to the resulting posterior odds ratio in Equation 14. We do this within the context of a simple normal sampling model under a conjugate prior.

To be specific, consider the choice between two models, $\mathcal{M}_1$ and $\mathcal{M}_0$, where y_i, a scalar variable of interest, is assumed to be generated as

$$y_i \mid \theta \sim \mathcal{N}(0, 1)$$

under $\mathcal{M}_1$ and

$$y_i \mid \theta \sim \mathcal{N}(\theta, 1)$$

under $\mathcal{M}_0$. For the unrestricted model $\mathcal{M}_0$, we employ the prior

$$\theta \mid \mathcal{M}_0 \sim \mathcal{N}(0, v_\theta).$$

As no unknowns are contained in the restricted $\mathcal{M}_1$, the marginal likelihood of Equation 15 (that is, the data density evaluated at the observed sample values) is immediate:

$$p(\boldsymbol{y}^o \mid \mathcal{M}_1) = (2\pi)^{-n/2} \exp\left(-\frac{1}{2}\sum_i (y_i^o)^2\right), \tag{16}$$

with y_i^o denoting the observed outcome for observation i, and $\boldsymbol{y}^o = [y_1^o\ y_2^o\ \ldots\ y_n^o]$ is the vector of outcomes. To evaluate the marginal likelihood under $\mathcal{M}_0$, we must calculate

$$p(\boldsymbol{y}^o \mid \mathcal{M}_0) = \int_{-\infty}^{\infty} \phi(\boldsymbol{y}^o; \theta \boldsymbol{\iota}_n, \boldsymbol{I}_n)\phi(\theta; 0, v_\theta)\, d\theta, \tag{17}$$

where $\phi(\boldsymbol{x}; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ denotes the normal density function for $\boldsymbol{x}$ with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$, $\boldsymbol{\iota}_n$ denotes an $n\times 1$ vector of ones, and $\boldsymbol{I}_n$ is the n-dimensional identity matrix.

To perform the integration in Equation 17 analytically, we must complete the square in θ, rearrange terms to produce a normal density for θ within the integrand, and keep track of the terms not involving θ. Doing so we obtain

$$p(\boldsymbol{y}^o \mid \mathcal{M}_0) = (2\pi)^{-n/2}(v_\theta + n^{-1})^{-1/2} n^{-1/2} \exp\left(-\frac{1}{2}\left[\left(\sum_{i=1}^{n} [y_i^o]^2\right) - \frac{n^2(\bar{y}_o)^2}{n + v_\theta^{-1}}\right]\right), \tag{18}$$

with $\bar{y}_o \equiv \frac{1}{n}\sum_{i=1}^{n} y_i^o$. Under equal prior odds, Equations 16 and 18 reveal that the posterior odds K_{01} reduce to

$$K_{01} = (v_\theta + n^{-1})^{-1/2} n^{-1/2} \exp\left(\left[\frac{v_\theta n}{1 + v_\theta n}\right]\frac{n(\bar{y}_o)^2}{2}\right). \tag{19}$$

Given this result, let us consider the testing exercise in practice. We have a fixed n and sample of data yielding $\bar{y}_o$. We now wish to consider the role of the prior in selecting between the restricted $\mathcal{M}_1$ and unrestricted $\mathcal{M}_0$.

For estimation purposes, many would let $v_\theta \to \infty$, as doing so would yield results that are close to the MLE and would approximate letting the data speak for themselves. The choice of such a prior, however, can have important and potentially unintended consequences for purposes of model comparison and selection. To see this, first note from Equation 19 that

$$\exp\left(\left[\frac{v_\theta n}{1 + v_\theta n}\right]\frac{n\bar{y}_o^2}{2}\right) \to \exp\left(\frac{n\bar{y}_o^2}{2}\right) \quad \text{as} \quad v_\theta \to \infty.$$

The directional impact of this contribution to the posterior odds ratio is sensible and intuitive, as larger values of $\bar{y}_o^2$ indicate a movement of the mean of the data away from zero, thus increasing the value of K_{01} and lending support for the unrestricted model.

As for the constant outside of the exponential kernel in Equation 19, however, we note

$$(v_\theta + n^{-1})^{-1/2} n^{-1/2} \to 0 \quad \text{as } v_\theta \to \infty$$

and thus $K_{01} \to 0$ as $v_\theta \to \infty$ In other words, it is possible to make the prior vague enough to favor the restricted model with (ex ante) probability one. This clearly reveals that seemingly default choices of conjugate priors with large variances, although reasonably innocuous for purposes of estimation, can have severe consequences for model comparison. As $v_\theta \to \infty$, the prior continually places less mass in a neighborhood of zero. As a result, any data information producing n and $\bar{y}^o$ can be interpreted as representing a relative movement toward zero, leading the researcher to support the restricted model. This rather perplexing result, showing that proper yet diffuse priors implicitly lend their support for the restricted model, is referred to as Bartlett's paradox (Bartlett 1957).

To illustrate Bartlett's paradox in practice and better understand the differential role of the prior in estimation and model comparison, consider the following simple linear regression application, using a sample of $n = 1{,}217$ observations from the National Longitudinal Survey of Youth:

$$y_i = \beta_0 + \beta_1 Ed_i + \varepsilon_i, \quad \varepsilon_i \,|\, \boldsymbol{Ed} \overset{iid}{\sim} N(0, \sigma^2).$$

In our model y refers to the log hourly wage received by individual i and Ed denotes education, or years of schooling completed. We also define

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \quad \boldsymbol{x}_i = [1 \;\; Ed_i], \quad \boldsymbol{X} = \begin{bmatrix} \boldsymbol{x}_1 \\ \boldsymbol{x}_2 \\ \vdots \\ \boldsymbol{x}_n \end{bmatrix}, \quad \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

and employ priors of the form

$$\beta \,|\, \sigma^2 \sim N(\mu, \sigma^2 V_\beta), \tag{20}$$

$$\sigma^2 \sim IG\left[\frac{\nu}{2}, 2(\nu\lambda)^{-1}\right], \tag{21}$$

with $\nu = 6$, $\lambda = 0.1333$, and $\mu = 0$ fixed throughout. Lastly, we consider two different priors: $\boldsymbol{V_\beta} = 10\boldsymbol{I}_2$ and $\boldsymbol{V_\beta} = 1.0{\times}10^{100}\boldsymbol{I}_2$.

Table 1 presents posterior statistics for this application under both priors. In terms of point estimates and marginal posterior distributions of the model parameters, there are essentially no differences in the results. With approximately 1,200 observations and two different priors that are both reasonably uninformative, the data information is dominant. Parameter posterior means under both priors are also nearly identical to conventional ordinary least squares (OLS) estimates, which are 1.18, 0.091, and 0.267 for the intercept, schooling coefficient, and variance parameter, respectively.

Let us now move beyond estimation and consider the question of model comparison. In terms of the economics of the problem, a key parameter of interest is β_1, commonly interpreted as the economic return to a year of education. The posterior mean of this parameter would imply that an added year of schooling will increase one's hourly wage by approximately 9.1% on average. The marginal posterior distribution of this parameter is also rather tightly concentrated around the posterior mean, with a comparably small posterior standard deviation of 0.007. In fact, when we employ a simulation-based

Table 1 Parameter posterior means, standard deviations (SDs), and marginal likelihoods from wage/education data set

Parameter	$\boldsymbol{V_\beta} = 10\,\boldsymbol{I}_2$		$\boldsymbol{V_\beta} = 1.0{\times}10^{100}\,\boldsymbol{I}_2$	
	Posterior mean	Posterior SD	Posterior mean	Posterior SD
β_0	1.17	(0.086)	1.18	(0.086)
β_1	0.091	(0.007)	0.091	(0.007)
σ^2	0.267	(0.011)	0.267	(0.011)
$\log K_{01}$	84.7		−29.9	

procedure to fit this model, the smallest of 25,000 simulations from this marginal posterior was 0.062. This provides seemingly overwhelming evidence that, within the context of this simple model, education matters in the production of postschooling earnings, as the marginal posterior is clearly bounded away from zero. Actually conducting a formal test of this claim by calculating the marginal likelihoods seems like an unnecessary exercise, as it is difficult to imagine supporting $\beta_1 = 0$ in light of this evidence. That said, let us suppose that two different researchers, one using $\boldsymbol{V}_\beta = 10\boldsymbol{I}_2$ and the other using $\boldsymbol{V}_\beta = 1.0\times10^{100}\boldsymbol{I}_2$, decided nonetheless to calculate the Bayes factor and posterior odds ratio.

Under the priors given in Equations 20 and 21, marginal data densities as in Equation 15 can be obtained analytically. Specifically, we obtain

$$\boldsymbol{y} \mid \boldsymbol{X}, \boldsymbol{V}_\beta, \lambda, \nu, \boldsymbol{\mu} \sim t_n\left[\boldsymbol{X}\boldsymbol{\mu}, \lambda(\boldsymbol{I}_n + \boldsymbol{X}\boldsymbol{V}_\beta\boldsymbol{X}'), \nu\right],$$

where t_n denotes an n-dimensional multivariate student-t density and is parameterized as in Koop et al. (2007, pp. 337–38). For a given $\boldsymbol{V}_\beta$, we can use this result to calculate the log marginal likelihood for both $\mathcal{M}_1$ (which drops education from $\mathbf{X}$) and $\mathcal{M}_0$ (which retains education). The difference between these two—offering the degree of support for $\mathcal{M}_0$ on the log scale—is reported in the last row of **Table 1**.

When setting $\boldsymbol{V}_\beta = 10\boldsymbol{I}_2$, we obtain $K_{01} \approx \exp(84.7) \approx 6.09\times10^{36}$. Thus our moderate prior strongly supports retention of the education variable. However, under the prior $\boldsymbol{V}_\beta = 1.0\times10^{100}\boldsymbol{I}_2$, we obtain $K_{01} \approx 1.03\times10^{-13}$, suggesting rather strongly that education should be dropped! This result offers an illustration of Bartlett's paradox: Although proper yet diffuse priors may let the data speak for themselves in terms of estimation, such priors can have significant consequences on model comparison, and indeed will tend to select the restricted model, even when seemingly overwhelming evidence is obtained to support the unrestricted variant of the model.

The results of this section suggest that considerable care must be taken with the specification of priors for purposes of model comparison. Although in many cases prior elicitation can be quite difficult, a useful suggestion is to think about implications of the prior $p(\boldsymbol{\theta})$ on the prior predictive distribution: $p(\boldsymbol{y}) = \int_{\boldsymbol{\Theta}} p(\boldsymbol{y} \mid \boldsymbol{\theta}) p(\boldsymbol{\theta}) d\boldsymbol{\theta}$. Although the researcher may have little guidance regarding reasonable values of $\boldsymbol{\theta}$, she is likely to have a much better sense for the location, spread, and other features of the outcome being modeled: Priors that give rise to prior predictive distributions with odd characteristics can be reformulated until $p(\boldsymbol{y})$ accords with the opinions of the researcher. Other, more automatic procedures for the selection of priors consist of unit-information-type priors (see, e.g., Kass & Wasserman 1995), which can result in simple approximations to the Bayes factor, like Bayesian information criterion, producing sensible results. Kass & Raftery (1995) also provide an extensive overview of Bayes factors and their application; DiCiccio et al. (1997) review the performances of a variety of different simulation-based strategies for calculating approximations to the Bayes factor; and Lopes & West (2004) compare these strategies to reversible jump MCMC when selecting the number of common factors in factor analysis. More recently, these automatic, unit-information priors have drawn the research attention of the fast-growing objective Bayes community (see, for instance, Berger 2006, Liang et al. 2008, Carvalho et al. 2010).[5]

[5]In fact, the Objective Bayesian Section of the International Society for Bayesian Analysis was founded in August 2010 and has James O. Berger as its first chair (**http://www.bayesian.org**).

3. BAYESIAN INFERENCE WITH FLEXIBLE SAMPLING MODELS

Whereas the previous section deals with issues of prior sensitivity, a second component of Bayesian analysis is the specification of a sampling density for the data or, viewed as a function of the parameters given the data, the likelihood function. As many examples in current and past Bayes textbooks (not unlike that provided in Section 2.5) assume normal sampling models for illustrative purposes, readers unfamiliar with Bayes may be left with the impression that the scope of Bayesian empirical work is limited to that spanned by the normal model, or perhaps the normal model plus a select few other special distributions.

In this section we seek to illustrate how other significantly more flexible representations of the likelihood can be handled in Bayesian empirical work. In the first part of this section we introduce flexibility in standard normal and/or linear models by altering their error/distributional structures. We begin with a stochastic volatility (SV) example in which student-t errors are indirectly introduced via a continuous-scale mixture of normals. We then consider the well-known class of a finite mixture of distributions, which is illustrated by a simple factor analysis whose common factors are modeled via a two-component mixture of normals. A DP mixture then is introduced as a nonparametric Bayes alternative to discrete mixtures that avoids issues related to fixing or estimating the number of mixture components.

3.1. Scale Mixture of Normals

A relatively simple way to generalize normality without losing many of its conveniences is to mix the normal model via its scale. In an example using a simple scalar scale mixture of normals (e.g., Andrews & Mallows 1974), a normally distributed random variable y is modeled as $N(\mu, \lambda\sigma^2)$, with the conditioning variable λ functioning as a mixture weight with assigned density $p(\lambda)$. Therefore, the marginal distribution of y is

$$p(y) = \int p(y \mid \lambda) p(\lambda) d\lambda.$$

Interestingly, many distributions may be obtained in this way. The most famous one is the student-t distribution with ν degrees of freedom, denoted t_ν, which emerges as the marginal density for y when $\lambda \sim IG(\nu/2, 2/\nu)$. Other known distributions that can be generated this way, upon the appropriate choice of mixing distribution $p(\lambda)$, include the double-exponential, logistic, and stable distributions.

> **Example 1 (stochastic volatility with student-*t* errors):** A popular model that can serve to illustrate the usefulness of such Gaussian extensions is the SV model. Univariate SV asset price dynamics result in the movements of an equity index S_t and its SV v_t via a continuous-time diffusion by a Brownian motion.[6] Because data arise in discrete time, it is natural to take an Euler discretization of the continuous-time diffusion. This then is commonly referred to as the stochastic volatility autoregressive model and is described by the following nonlinear dynamic model (West & Harrison 1997):

[6]More explicitly, $d \log S_t = \mu dt + \sqrt{v_t}\, dB_t^P$ and $d \log v_t = \kappa(\gamma - \log v_t)dt + \sigma dB_t^V$, where the parameters governing the volatility evolution are $(\mu, \kappa, \gamma, \sigma)$, and Brownian motions (B_t^P, B_t^V) are possibly correlated (Rosenberg 1972, Taylor 1986, Hull & White 1987, Ghysels et al. 1996, Johannes & Polson 2010).

$$y_t = \exp\{x_t/2\}\varepsilon_t, \tag{22}$$

$$x_t = \beta_0 + \beta_1 x_{t-1} + \sigma\eta_t, \tag{23}$$

where y_t and x_t are log returns and log variances, respectively; ε_t and η_t are i.i.d. standard normal errors; and the initial log-volatility state x_0 follows $N(m_0, C_0)$, for known m_0 and C_0. The SV model is completed with a conjugate prior distribution for $(\beta_0, \beta_1, \sigma^2)$, i.e., $p(\beta_0, \beta_1|\sigma^2)p(\sigma^2)$, where $(\beta|\sigma^2) \sim N(b_0, \sigma^2 B_0)$ and $\sigma^2 \sim IG[v_0/2, 2/(v_0\sigma_0^2)]$, for known b_0, B_0, v_0, and σ_0^2.

The prior specification for (x_0, β, σ^2) can be treated as a model choice or prior robustness problem.[7] More importantly, however, the model assumes normality of the ε's, which may not be adequate in practice. A model that is robust to extreme observations or outliers in the returns equation can be achieved, for instance, by entertaining fat-tailed distributions for ε_t in Equation 22. One possible way is to consider a continuous-scale mixture of normals (Carlin & Polson 1991, Geweke 1993, Jacquier et al. 2004, Lopes & Polson 2010c):

$$y_t = \exp\{x_t/2\}\varepsilon_t, \tag{24}$$

$$x_t = \beta_0 + \beta_1 x_{t-1} + \sigma\eta_t, \tag{25}$$

$$\varepsilon_t = \sqrt{\lambda_t}\, z_t, \tag{26}$$

$$\lambda_t \sim IG(v/2, 2/v), \tag{27}$$

where z_t's are i.i.d. standard normal so that, after integrating out λ_t, ε_t is distributed as $t_v(0,1)$, a standard student-t distribution with v degrees of freedom. We call this the SV-t_v model, which can also accommodate a wide range of kurtosis behavior.[8]

For numerical illustration, we fit three SV models to the $n = 391$ daily log returns on the S&P 500 Index corresponding to business days of 2009 and 2010 up to July 22.[9] The models are the normal SV model, denoted here by

[7]An alternative specification for x_0 assumes that $(x_0|\beta_0, \beta_1, \sigma^2) \sim N[\beta_0/(1-\beta_1), \sigma^2/(1-\beta_1^2)]$ with $|\beta_1| < 1$. The researcher can apply the resampling tool described above to study the sensitivity of the posterior distribution of $(\beta_0, \beta_1, \sigma^2, x_1, \ldots, x_n)$ to the choice of the prior for x_0. In our experience, the effect of the prior is negligible in most SV models unless n is relatively small and/or β_1 is fairly close to one, i.e., a unit root problem (Kalayloglu & Ghosh 2009). Yet another parameterization moves β_0 to the observation equation and centers the x_t's. This only marginally affects posterior inference in most cases while creating unnecessary computational burden. A common alternative for (β, σ^2) considers β and σ^2 independent a priori and can be easily implemented with negligible additional computational cost. Also negligible, in our experience, is the effect of these modifications on the posterior distributions of the model parameters.

[8]Additional contributions to the theme are Steel (1998), Chib et al. (2002), Omori et al. (2007), Asai (2009), Nakajima & Omori (2009), and Abanto-Valle et al. (2010). Lopes & Polson (2010b) provide a sequential Monte Carlo scheme and analysis of the credit crises of 2007–2008 by comparing standard SV models to the ones with stochastic jump components. Lopes & Polson (2010a) provide a recent and detailed review of Bayesian inference for univariate and multivariate SV models.

[9]Data were collected from **http://finance.yahoo.com**. Log return y_t is computed as log p_t/p_{t-1}, where p_t is the close price adjusted for dividends and yields.

Table 2 Posterior summary for model parameters

Parameter	SV-t_∞ ($\mathcal{M}_0$)		SV-t_{10} model ($\mathcal{M}_1$)		SV-t_1 model ($\mathcal{M}_2$)	
	Median	95% CI	Median	95% CI	Median	95% CI
β_0	−0.242	(−0.479, −0.046)	−0.242	(−0.485, −0.041)	−0.247	(−0.517, −0.025)
β_1	0.975	(0.963, 0.983)	0.976	(0.963, 0.984)	0.976	(0.963, 0.985)
σ	0.453	(0.343, 0.604)	0.457	(0.351, 0.614)	0.496	(0.375, 0.666)

Abbreviations: CI, credibility interval; SV, stochastic volatility.

SV-t_∞ ($\mathcal{M}_0$); the SV-t_{10} model ($\mathcal{M}_1$); and SV-t_1 model ($\mathcal{M}_2$). The priors employed for these models are somewhat informative, and hyperparameters were selected to be consistent with results obtained in previous SV studies.[10] Posterior inference for $(\beta_0, \beta_1, \sigma)$ appears in **Table 2**, whereas estimates of the standard deviation of log returns are presented in **Figure 1**. Posterior summaries are fairly similar across models, whereas the SV-t_1 model clearly is more conservative than the other two models when estimating the standard deviations.

A natural question to ask is, of the three models, which is the best? The Bayes factor of $\mathcal{M}_0$ versus $\mathcal{M}_1$ (or $\mathcal{M}_2$) can be approximated by the Savage-Dickey density ratio (Verdinelli & Wasserman 1995, Marin & Robert 2010) by noting that (*a*) $\mathcal{M}_0$ is nested in $\mathcal{M}_1$ (or $\mathcal{M}_2$) when, for Equation 27, $\lambda_t = 1$ for all t, and (*b*) the prior distributions for $(\beta_0, \beta_1, \sigma^2)$ under all three models are the same. In this case

$$B_{01} = \frac{p(\lambda_1 = 1, \ldots, \lambda_n = 1 \,|\, \mathbf{y}, \mathcal{M}_1)}{p(\lambda_1 = 1, \ldots, \lambda_n = 1 \,|\, \mathcal{M}_1)} \approx \frac{\prod_{t=1}^{n} \left\{ \frac{1}{N} \sum_{i=1}^{N} f_{IG}\left(1; \frac{v+1}{2}, \frac{v + y_t^2 e^{-x_t^{(i)}}}{2}\right) \right\}}{\left\{ f_{IG}\left(1; \frac{v}{2}, \frac{v}{2}\right) \right\}^n},$$

where $x_t^{(i)} (i = 1, \ldots, N)$ are posterior draws for log volatilities. We found that $\log B_{01}$ and $\log B_{02}$ are approximately −160.1 and −33.0, respectively. Hence there is strong evidence supporting t_{10} or t_1 over the normal model and supporting t_1 over t_{10}. Clearly, for this application, the extension to a more general sampling model is strongly supported by the data. **Table 3** presents the logarithm of the Bayes factor of the normal SV model versus the SV-t_v model for a few different values of v. As expected, when v goes to infinity, which translates into t_v approximating the normal, the logarithm of the Bayes factor goes to zero.

[10]Prior hyperparameters for all models are $b_0 = (-0.2, 0.98)'$, $B = \text{diag}(0.15, 0.001)$, $v_0 = 10$, $\sigma_0^2 = 0.22$, $m_0 = 0$, and $C_0 = 1$. All MCMC algorithms started at $\left[\beta_0^{(0)}, \beta_1^{(0)}, \sigma^{2(0)}\right] = (-0.2, 0.98, 0.22)$ and were run for $M_0 = 5{,}000$ iterations followed by $M = 5{,}000$ kept for posterior inference.

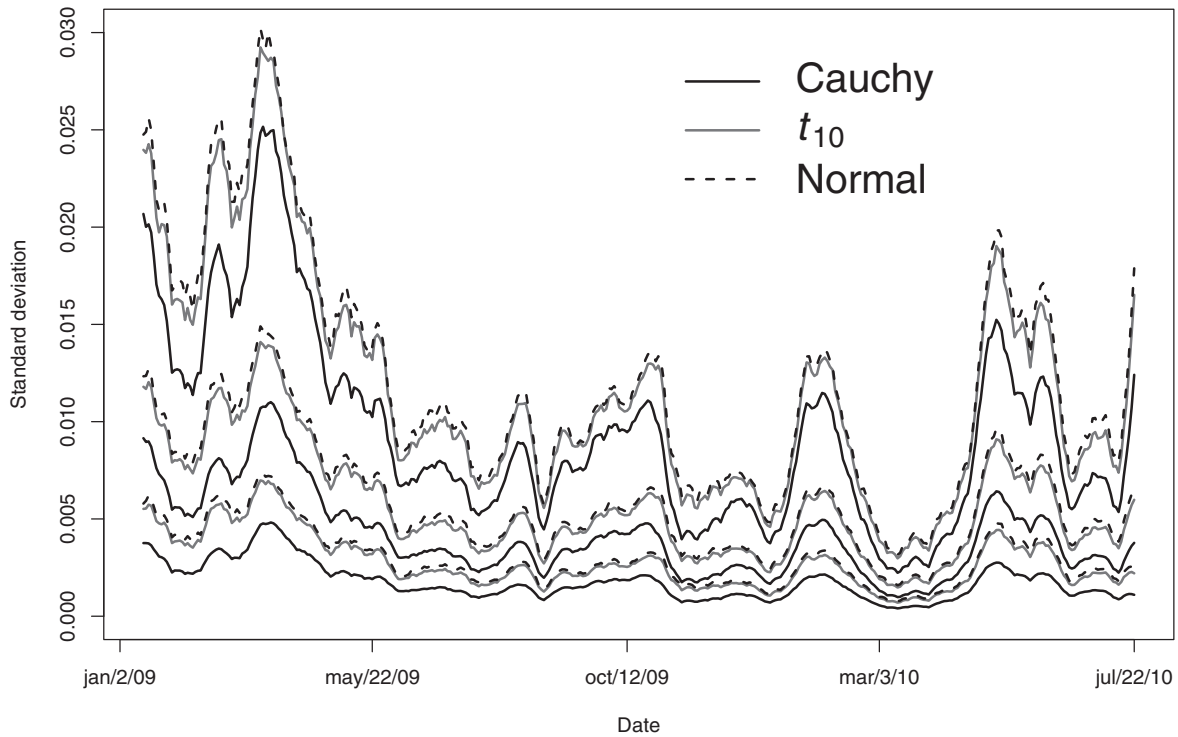

Figure 1

Stochastic volatility. Posterior median and 95% credibility interval for the standard deviation of log returns y_t, $\exp\{x_t/2\}$, for three stochastic volatility models: the normal model ($\mathcal{M}_0$, *dashed line*), t_{10} model ($\mathcal{M}_1$, *gray line*), and Cauchy model ($\mathcal{M}_2$, *black line*).

Finally, to compute posterior model probabilities for all the competing models entertained in **Table 3**, we need to compute all Bayes factors. In addition, we must introduce prior model probabilities. For simplicity, we assume that each model is equally likely a priori or that $Pr(\mathcal{M}_i) = 1/M$, where M is the number of competing models. The log Bayes factor of model $\mathcal{M}_i$ versus $\mathcal{M}_j$ does not need to be recomputed for each pair under consideration, but instead can be obtained as $\log B_{ij} = \log B_{j0} - \log B_{i0}$. Then, it is easy to show that $Pr(\mathcal{M}_i \mid \text{data}) = 1/\sum_k B_{ki}$, for $i = 1, \ldots, M$. Based on the results of **Table 3**, we can see that the posterior probability associated with the Cauchy model is virtually one, which again clearly emphasizes the value in extending the standard Gaussian sampling model for this application.

3.2. Finite Mixture of Distributions

General and flexible classes of distributions to model heterogeneous data can be provided by a discrete mixture of distributions or, more generally, mixture models (Dalal & Hall 1983). In this section we briefly review finite mixtures with an emphasis on the issue of labeling or classifying observations to mixture components and on outlining an MCMC scheme for posterior inference. More details on modeling and computation in mixture

Table 3 Weight of evidence

i	ν	$\log B_{0i}$
1	1	−160.107480
2	10	−33.006898
3	100	−5.835605
4	1,000	−0.765586
5	10,000	−0.078399
6	100,000	−0.008314
7	1,000,000	−0.000081

models are abundant in Titterington et al. (1985), Diebolt & Robert (1994), Escobar & West (1995), McLachlan & Peel (2000), and Frühwirth-Schnatter (2006), among others.

Let the probability density function or the probability distribution of y be

$$p(y \mid \gamma) = \sum_{j=1}^{k} \pi_j p_j(y \mid \theta_j),$$

where $\gamma = (\theta_1, \ldots, \theta_k, \pi_1, \ldots, \pi_k)$, $\pi_j > 0$ for $j = 1, \ldots, k$, and $\sum_{j=1}^{k} \pi_j = 1$. The probability density function p_j is parameterized by θ_j. The vectors $\theta_1, \ldots, \theta_k$ may share a common component, such as in a mixture of two normal components with common variance. We assume for the moment that the number of mixture components k is fixed and known. Estimation of the number of mixture components will be considered below in Section 3.3.

3.2.1. Label switching. A potential problem with the use of finite mixtures in practice is one of label switching. If the main inferential goal is to identify and interpret the mixture components and/or clustering of the observations, then this becomes an issue and should be carefully handled. The problem of label switching refers to the fact that the data density is unaltered upon permuting the subscripts: It is easy to see that $p(y \mid \gamma) = p(y \mid \tilde{\gamma})$, where $\tilde{\gamma} = (\theta_{j_1}, \ldots, \theta_{j_k}, \pi_{j_1}, \ldots, \pi_{j_k})$ and $j_1, \ldots, j_k$ any permutation of $1, \ldots, k$. In this case the simulations produced from one particular mixture component need not correspond to any meaningful quantity, as nothing has been put in place to individually identify the mixture components and thereby cement the interpretation of the component parameters.

Conversely, if the main goals of the analysis involve posterior prediction, density estimation, or posterior inference for a quantity common across components, then label switching is no longer a problem. To illustrate, we note that conditional on observations $\boldsymbol{y}$, the posterior predictive density of y_{n+1} is

$$p(y_{n+1} \mid \boldsymbol{y}) = \int p(y_{n+1} \mid \gamma, \boldsymbol{y}) p(\gamma \mid \boldsymbol{y}) d\gamma = \int p(y_{n+1} \mid \gamma) p(\gamma \mid \boldsymbol{y}) d\gamma,$$

where $p(\gamma \mid \boldsymbol{y})$ is the posterior distribution of γ, and the second equality follows from the assumption of conditionally independent observations. Therefore, the configuration of γ is of no interest for purposes of posterior prediction. For more discussion on the importance

(or irrelevance) of label switching, readers are referred to Stephens (2000), Jasra et al. (2005), and particularly Geweke (2007).

3.2.2. Posterior inference via Markov chain Monte Carlo. Modern Bayesian inference in a mixture of distributions is mostly done via a data-augmentation argument. A fictitious group classifier/indicator, e.g., z_i, is introduced for every observation y_i. Specifically, observation i is classified as belonging to group j when $z_i = j$. In this case, the conditional posterior distribution of γ (which also conditions on the latent data z) can be written as

$$p(\gamma \mid \boldsymbol{y}, \boldsymbol{z}) \propto p(\gamma) \prod_{j=1}^{k} \prod_{i:z_i=j} p_j(y_i \mid \theta_j).$$

If, additionally, $p(\gamma)$ can be decomposed into $p(\gamma) = p(\pi)\prod_{j=1}^{n} p(\theta_j)$, then it follows that the posterior distribution of γ is also decomposable:

$$p(\gamma \mid \boldsymbol{y}, \boldsymbol{z}) \propto p(\pi) \prod_{j=1}^{k} \prod_{\{i:z_i=j\}} p(\theta_j) p_j(y_i \mid \theta_j).$$

This separation leads directly to a Gibbs sampler with the following full conditional distributions: (*a*) $p(\theta_j \mid \theta_{-j}, \pi, \boldsymbol{y}, \boldsymbol{z}) \equiv p(\theta_j \mid \boldsymbol{y}, \boldsymbol{z}) \propto \prod_{i:z_i=j} p(\theta_j) p_j(y_i \mid \theta_j)$, where $\theta_{-1} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_k)$; (*b*) $p(\pi \mid \boldsymbol{y}, \boldsymbol{z}) \equiv p(\pi \mid \boldsymbol{z}) \propto p(\pi)\prod_{j=1}^{n} \pi_j^{n_j}$, where n_j is the number of observations classified in group j; and (*c*) $p(\boldsymbol{z} \mid \boldsymbol{y}, \gamma) \propto \prod_{i=1}^{n} p(z_i) p(y_i \mid \theta_i, z_i)$. By letting the prior distribution of π be a Dirichlet(α), it follows that $(\pi \mid \boldsymbol{y}, \boldsymbol{z})$ is also Dirichlet($\alpha + \boldsymbol{n}$) for $\boldsymbol{n} = (n_1, \ldots, n_k)$. Finally, for $i = 1, \ldots, n$, $(z_i \mid y_i, \gamma) \sim \{1, \ldots, k\}$, with $Pr(z_i = j \mid y_i, \gamma) = \pi_j p(y_i \mid \theta_j)/c_i$ and $c_i = \sum_{l=1}^{k} \pi_l p(y_i \mid \theta_l)$. For model details on Bayesian inference when the number of mixture components k is unknown, readers are referred to Richardson & Green (1997), Lopes et al. (2003), and Dellaportas & Papageorgiou (2006), to name a few.

> **Example 2 (factor analysis with a mixture of common factors):** Factor models represent one of the most used (and useful) statistical techniques and are applied mainly in data reduction and the identification of underlying data structures. In the standard normal factor model, p continuous measurements in y_i for individual (or unit) i are conditionally independent given k common factors f_i:
>
> $$y_i \mid f_i \sim \mathcal{N}(\beta f_i, \Sigma),$$
>
> where β is the $(p \times k)$ matrix of factor loadings, and Σ is the diagonal matrix of idiosyncratic variances. A common identification constraint is to impose that β is block lower triangular with ones in the main diagonal (see Lopes 2003 for an annotated bibliography on factor models). The model is then complete given a prior specification for the common factors. The norm is to model f_i by a k-variate normal with zero mean vector and diagonal covariance matrix H (see Lopes & West 2004 and Frühwirth-Schnatter & Lopes 2010 for simulation-based Bayesian inference for high-dimensional factor models when the number of common factors k is unknown). Conti et al. (2011) use the latter to construct economically justified aggregates in a health economics application.
>
> For illustration, we generate data and then fit a simplistic factor model.[11] We obtain $n = 1{,}000$ observations over $p = 3$ measurements and consider a common

[11] A more general version of this model was used in Carneiro et al. (2003) to study returns to schooling and the effects of uncertainty on college choice.

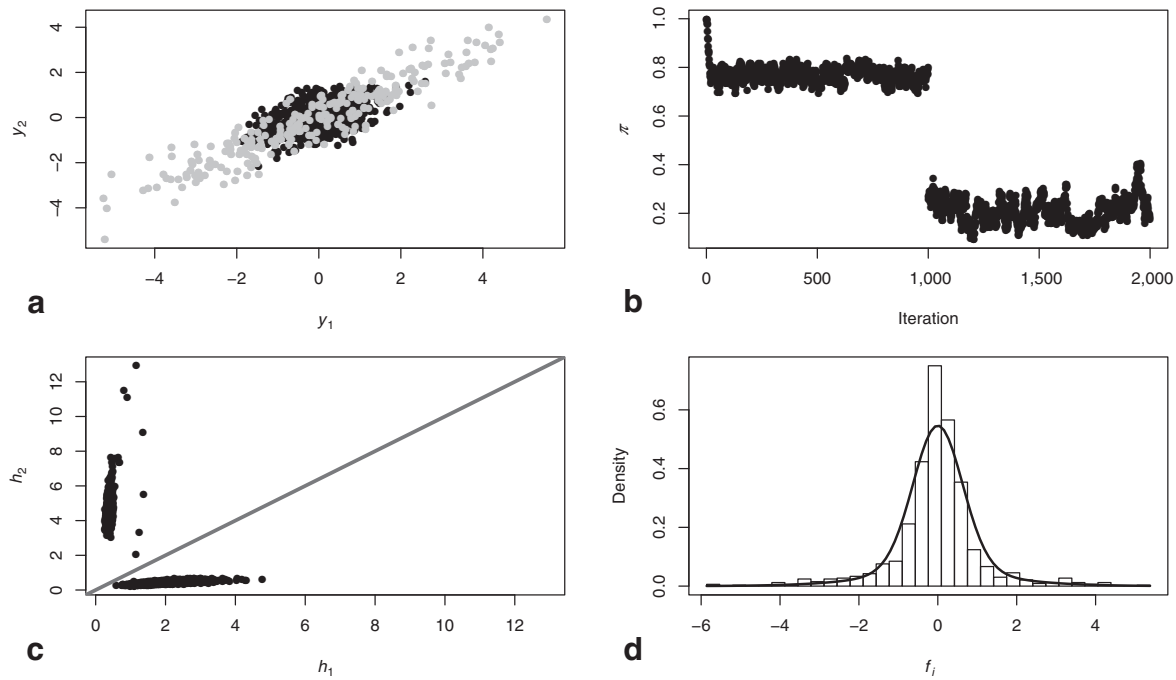

Figure 2

Factor mixture. (*a*) Measurements y_1 and y_2. (*b*) Markov chain Monte Carlo (MCMC) trace plot for draws of parameter π. (*c*) MCMC draws for (h_1,h_2). (*d*) True common factors (histogram) and posterior predictive of common factor.

factor ($k = 1$) that follows a two-component mixture of normals. Specifically, we consider $y_i \,|\, f_i \sim \mathcal{N}(\beta f_i, \Sigma)$ and $f_i \sim \pi\mathcal{N}(0, h_1) + (1 - \pi)\mathcal{N}(0, h_2)$, where $\beta = (1, 0.8, 0.25)'$, $\Sigma = 0.2I_3$, $\pi = 0.75$, $h_1 = 0.25$, and $h_2 = 4.0$. The simulated pair (y_1, y_2) appears in **Figure 2*a***. The same model is fit to the data via MCMC with independent priors for the components of $\theta = (\sigma_1^2, \sigma_2^2, \sigma_3^2, \beta_2, \beta_3, h_1, h_2, \pi)$: $\sigma_i^2 \sim IG(3, 1/0.4)(i = 1, 2, 3)$, $\beta_j \sim N(0,10)(j = 2, 3)$, $h_1 \sim IG(3, 2)$, $h_2 \sim IG(3, 1/8)$, and $\pi \sim Beta(1, 1)$.

MCMC initial values are based on classical estimation of the model when assuming normal factor scores. **Figure 2*b*,*c*** illustrates the label-switching problem, as **Figure 2*b***, for example, illustrates the switch in the interpretation of the first and second components around iteration 1,000. This problem is irrelevant, however, when assessing the posterior predictive distribution of the common factor for a new observation (**Figure 2*d***) because $p(f_{n+1} \,|\, \boldsymbol{y}) = \int \{\pi N(0, h_1) + (1 - \pi)N(0, h_2)\} p(\pi, h_1, h_2 \,|\, \boldsymbol{y}) d\pi dh_1 dh_2$.

3.3 Dirichlet Process Mixture

Since the early work of Ferguson (1974) and Antoniak (1974), discrete nonparametric mixture models have emerged as a dominant modeling tool for Bayesian nonparametric density estimation. Unlike our previous discussion of finite mixture models, in which the number of mixture components was regarded as fixed and chosen by the researcher,

nonparametric mixtures regard the number of components as random and enable data-based estimation of the number of such components. As the DP mixture is the most commonly used nonparametric prior for such random mixture models, we now discuss the basics of posterior inference with DP mixtures.

The DP characterizes a prior over probability distributions. Here we concentrate on the context in which observations $y_1, \ldots, y_n$ are independently drawn from some unknown distribution:

$$\begin{aligned} y_i \mid \theta_i &\sim F(y_i \mid \theta_i), \\ \theta_i \mid G &\sim G, \\ G \mid \alpha, G_0 &\sim DP(\alpha, G_0). \end{aligned}$$

Therefore, the DP is determined by two parameters: a distribution function G_0 defining the location of the DP prior and a precision parameter $\alpha > 0$. For any partition $B_1, \ldots, B_m$ on the space of support for G_0, the vector of probabilities $[G(B_1), \ldots, G(B_m)]$ has a Dirichlet distribution with parameter $[\alpha G_0(B_1), \ldots, \alpha G_m(B_m)]$. The precision parameter α can be thought of as a measurement of the concentration of the prior around G_0. In many circumstances, G_0 is the usual parametric prior specification for the problem at hand. The DP adds another stage in a hierarchical model that has as a baseline a common parametric model. When α is large, samples from G will be close to G_0. When α is small, samples of G will likely concentrate on a few so-called atoms, so the DP will mimic the behavior of finite mixture models.

3.3.1. Stick-breaking argument. An important result, known as the stick-breaking argument, is useful for sampling DP realizations. A random distribution G generated from DP (α, G_0) is almost surely of the form

$$dG(\cdot) = \sum_{l=1}^{\infty} \pi_l \, \delta_{\theta_l}(\cdot), \tag{28}$$

where δ_{θ_l} is a point mass located at θ_l, $\pi_l = (1 - \sum_{j=1}^{l-1} \pi_j) v_l \, (l = 1, 2, \ldots)$, θ_l represents i.i.d. draws from G_0, and v_l represents i.i.d. draws from $Beta(1, \alpha)$ (Perman et al. 1992). The discreteness of DP realizations is explicit in this definition.

3.3.2. Conjugate family of priors. The DP provides a conjugate family of priors over distributions that are closed under posterior updates given observations. This can be seen by integrating out G to obtain the prior distribution of the θ_i's in terms of successive conditional distributions (Blackwell & MacQueen 1973):

$$\theta_i \mid \theta_1, \ldots, \theta_{i-1} \sim \left(\frac{\alpha}{i - 1 + \alpha} \right) G_0 + \left(\frac{i - 1}{i - 1 + \alpha} \right) \frac{\sum_{j=1}^{i-1} \delta_{\theta_j}}{i - 1}.$$

Then it follows directly that the posterior distribution over G is also a DP with concentration $\alpha + n$ and base measure

$$\left(\frac{\alpha}{\alpha + n} \right) G_0 + \left(\frac{n}{\alpha + n} \right) \frac{\sum_{i=1}^{n} \delta_{\theta_i}}{n};$$

i.e., the base measure is a weighted average between the prior base distribution G_0 and the empirical distribution $\frac{1}{n} \sum_{i=1}^{n} \delta_{\theta_i}$.

3.3.3. Posterior inference. When the baseline distribution G_0 and the model likelihood, denoted here by $f(y_i \mid \theta_i)$, conjugate, the simplest and most direct approach to sampling from the DP model is to iterate draws from the full conditional distributions

$$\theta_i \mid \theta_{-i} \sim r_i H_i + \sum_{j \neq i} q_{ij} \delta_{\theta_j}, \tag{29}$$

where H_i is the posterior distribution of θ_i based on the baseline distribution G_0 and y_i, $q_{ij} = bf(y_i \mid \theta_j)$, $r_i = b\alpha \int f(y_i \mid \theta) dG_0(\theta)$, and b is such that $r_i + \sum_{j \neq i} q_{ij} = 1$. The posterior predictive for a new observation y_{n+1} can be approximated by Monte Carlo using the identity

$$p(y_{n+1} \mid \boldsymbol{y}) = \int f(\boldsymbol{y}_{n+1} \mid \theta) dG(\theta \mid \boldsymbol{y}) \approx \frac{1}{M} \sum_{m=1}^{M} f(y_{n+1} \mid \theta^{(m)}).$$

The following example provides an illustration of this approximation.

The standard approach to inference is to use a collapsed Gibbs sampler (Escobar & West 1995). Extension to the case of conditionally nonconjugate kernel models is described by Bush & MacEachern (1996), and Neal (2000) provides an overview of more state-of-the-art versions of the algorithm. Furthermore, Ishwaran & James (2001) describe sampling for truncated G approximations that can be utilized whenever it is not possible to marginalize over unallocated mixture components. Sequential Monte Carlo methods applied to general mixture models are discussed by Carvalho et al. (2009, and references therein). They build on the recent particle learning framework for sequential Bayesian computation of Lopes et al. (2010) and Carvalho et al. (2010) (see also Lopes & Tsay 2010 for a recent overview of sequential Monte Carlo techniques in financial time-series analysis). For additional references, the reader is directed to classical papers and (edited) books, including Dey et al. (1998), Walker et al. (1999), Ghosh & Ramamoorthi (2003), Müller & Quintana (2004), and Hjort et al. (2010).

> **Example 3 (linear regression with non-Gaussian errors):** We revisit the National Longitudinal Survey of Youth data from Section 2.5 in which y_i is the hourly wage (in logs) received by individual i and x_i is his years of schooling completed, i.e., $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$, for $i = 1, \ldots, n$. We entertain here three alternative model structures for the i.i.d. error terms: the baseline model $\mathcal{M}_0$, where $\varepsilon_i \sim N(0, \sigma^2)$; the student-$t$ model $\mathcal{M}_1$, where $\varepsilon_i \sim t_\nu(0, \sigma^2)$; and the DP model $\mathcal{M}_2$, where $\varepsilon_i \sim N(0, \sigma_i^2)$, $\sigma_i^2 \sim G$, $G \sim DP(\alpha, G_0)$, $G_0 \equiv IG[\eta/2, 2/(\eta\lambda)]$, and α, η, and λ are fixed. For the DP model $\mathcal{M}_2$, σ_i^2 plays the role of θ_i, and ε_i (conditional on β_0 and β_1) plays the role of y_i in Equation 29. In addition, $F(\cdot \mid \sigma_i^2) \equiv N(0, \sigma_i^2)$, and $H_i \equiv IG[(\eta + 1)/2, 2/(\eta\lambda + \varepsilon_i^2)]$. It is easy to see that $r_i = b\alpha t_\eta(\varepsilon_i; 0, \lambda)$ and $q_{ij} = bf_N(\varepsilon_i; 0, \sigma_j^2)$ with b following directly.[12]
>
> We run two MCMC chains with both starting at the OLS estimates for β_0, β_1, and σ^2, with the first chain starting at $\nu^{(0)} = 1$ and the second chain starting at $\nu^{(0)} = 100$. The chains are warmed up for $M_0 = 1{,}000$ iterations, and every

[12]The prior for $\beta \mid \sigma^2$ is $N(0, \sigma^2 10 I_2)$, and for σ^2 it is $IG(3, 2.5)$, the same in Section 2.5 for the normal model. The prior for ν is $G(1, \nu_0)$ with $\nu_0 = 25$ as suggested by Geweke (1993). This prior specification allocates substantial prior probability on values of ν below 10 (fat tails) as well as above 40 (normality).

hundredth draw of the following 100,000 draws was kept for posterior summarization, producing a total of $M = 2{,}000$ draws. **Figure 3** suggests strong agreement between both the normal and student-t model. This agreement is corroborated by the 95% credibility interval for ν, namely $(24.5, 102.5)$, which is away from a heavy-tail behavior. The same interval has prior probability of roughly 35%.

The DP process enters here to illustrate how Bayesian density estimation can be performed in this example. Again, for simplicity, we assume that β_0 and β_1 are known, so our observations are the ε_i's. In this case, the posterior predictive for the error term can be approximated by

$$p(\varepsilon_{n+1} \mid \varepsilon) \approx \frac{1}{M} \sum_{m=1}^{M} f[\varepsilon_{n+1} \mid \Theta^{(m)}],$$

where $\Theta^{(m)} = (\sigma_1^2, \ldots, \sigma_n^2)^{(m)}$, $m = 1, \ldots, M$ are MCMC draws from the DP model (as described above), and

$$f(\varepsilon_{n+1} \mid \Theta) = \frac{\alpha}{\alpha + n} f_T(\varepsilon_{n+1}; \eta, 0, \lambda) + \frac{1}{\alpha + n} \sum_{i=1}^{n} f_N(\varepsilon_{n+1}; 0, \sigma_i^2).$$

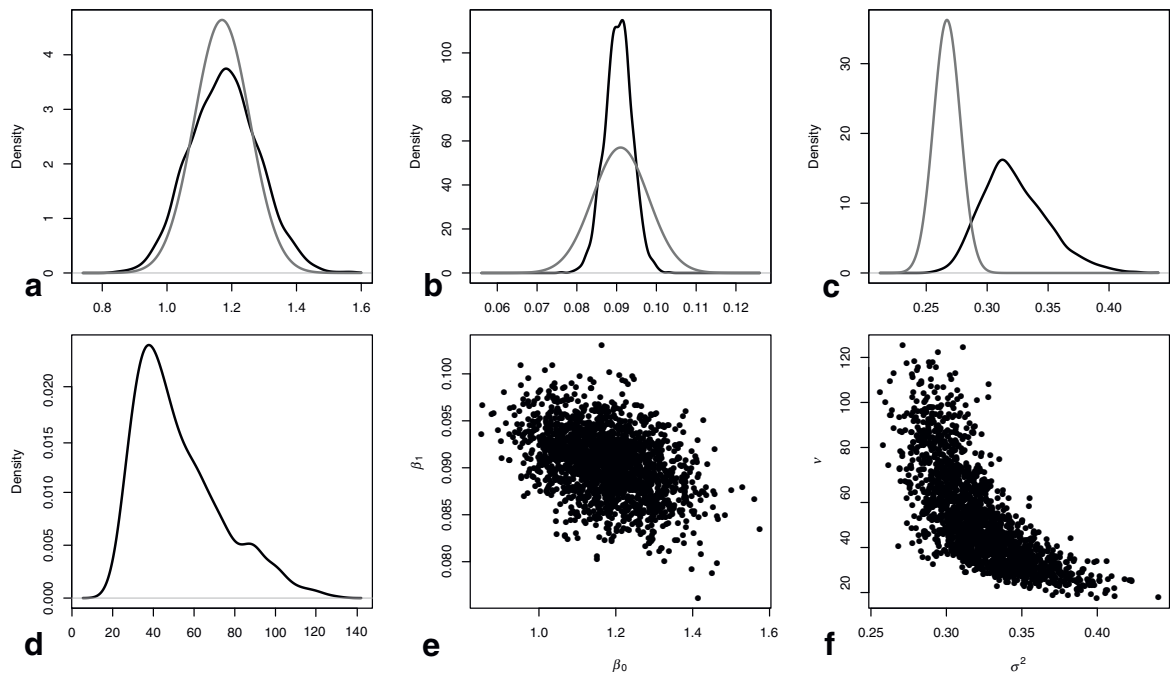

Figure 3

Student-t regression. Panels *a–c* are kernel-based approximations to $p(\beta_0 \mid \boldsymbol{y}, \mathcal{M}_i)$, $p(\beta_1 \mid \boldsymbol{y}, \mathcal{M}_i)$, and $p(\sigma^2 \mid \boldsymbol{y}, \mathcal{M}_i)$, respectively, under the normal model $\mathcal{M}_0$ (*gray lines*) and the student-t model $\mathcal{M}_1$ (*black lines*), and panels *d–f* approximate $p(\nu \mid \boldsymbol{y}, \mathcal{M}_1)$, $p(\beta_0, \beta_1 \mid \boldsymbol{y}, \mathcal{M}_1)$, and $p(\sigma^2, \nu \mid \boldsymbol{y}, \mathcal{M}_1)$, respectively.

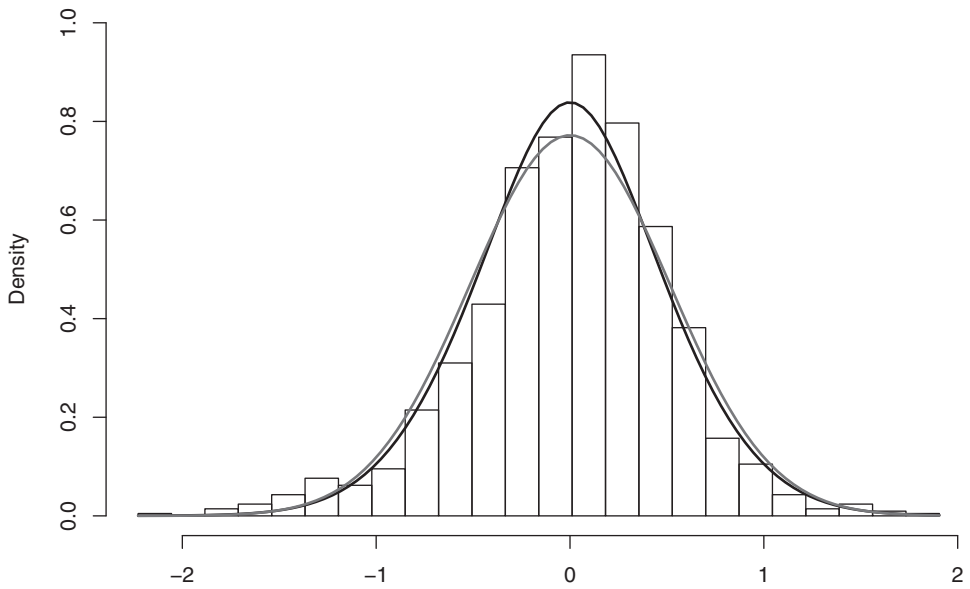

Figure 4

Dirichlet process (DP). Shown is $p(\varepsilon_{n+1} \mid \varepsilon)$ along with the histogram of ε. The DP model is represented as the black line, and the normal approximation is the gray line. The concentration parameter for the DP is set to $\alpha = 1$.

Figure 4 exhibits the approximate posterior predictive for the error term in the above linear regression when β_0 and β_1 are known and equal to the OLS estimates. Similar to the normal model and the student-t model, $\mathcal{M}_0$ and $\mathcal{M}_1$, respectively, the DP model for $\boldsymbol{\varepsilon}$ is fairly close to normality.

4. CONCLUSION

Above we review several results and at times offer a few personal thoughts on the sensitivity of Bayes results to the prior and likelihood. With respect to the prior, we argue that all analyses—Bayesian or non-Bayesian—incorporate some amount of prior information, so this issue is not really specific to the Bayesian approach. We also argue that in many settings, Bayes and frequentist estimators are asymptotically equivalent, and thus, under this widely adopted method for inference, concerns regarding the prior are largely unfounded. Finally, we discuss the significant role for the prior in terms of model selection and comparison and describe a simple simulation-based method to quantify how reported posterior means will change as the prior changes. With respect to the form of the likelihood, we also review a number of flexible yet computationally tractable representations of the sampling model via scale and finite mixtures of Gaussian distributions as well as models based on DPs.

Although we just scratch the surface in this article, we invite the reader to seek more information in a number of recent Bayesian texts, including Poirier (1995), Carlin & Louis (2000), Koop (2003), Gelman et al. (2004), Lancaster (2004), Geweke (2005), Gamerman & Lopes (2006), Koop et al. (2007), and Greenberg (2008). Furthermore, the volume edited by Insua & Ruggeri (2000) contains a number of articles, at a more technical level than that presented here, on Bayesian robustness and related issues.

DISCLOSURE STATEMENT

The authors are not aware of any affiliations, memberships, funding, or financial holdings that might be perceived as affecting the objectivity of this review.

LITERATURE CITED

Abanto-Valle CA, Bandyopadhyay D, Lachos VH, Enriquez I. 2010. Robust Bayesian analysis of heavy-tailed stochastic volatility models using scale mixtures of normal distributions. *Comput. Stat. Data Anal.* 54:2883–98

Andrews DF, Mallows CL. 1974. Scale mixtures of normality. *J. R. Stat. Soc. B* 36:99–102

Antoniak C. 1974. Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *Ann. Stat.* 2:1152–74

Asai M. 2009. Bayesian analysis of stochastic volatility models with mixture-of-normal distributions. *Math. Comput. Simul.* 79:2579–96

Bartlett MS. 1957. Comment on D.V. Lindley's statistical paradox. *Biometrika* 44:533–34

Berger JO. 2006. The case for objective Bayesian analysis. *Bayesian Anal.* 1:385–402

Bernardo JM, Smith AFM. 2000. *Bayesian Theory.* Chichester, UK: Wiley & Sons. 586 pp.

Blackwell D, MacQueen JB. 1973. Ferguson distributions via Polya urn schemes. *Ann. Stat.* 1:353–55

Bush C, MacEachern S. 1996. A semiparametric Bayesian model for randomized block designs. *Biometrika* 83:275–85

Carlin BP, Louis TA. 2000. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: CRC. 419 pp.

Carlin BP, Polson NG. 1991. Inference for non-conjugate Bayesian models using the Gibbs sampler. *Can. J. Stat.* 19:399–405

Carneiro P, Hansen KT, Heckman JJ. 2003. Application to the returns to schooling and measurement of the effects of uncertainty on college choice. *Int. Econ. Rev.* 44:361–422

Carvalho CM, Johannes MS, Lopes HF, Polson NG. 2010. Particle learning and smoothing. *Stat. Sci.* 25:88–106

Carvalho CM, Lopes HF, Polson NG, Taddy M. 2009. Particle learning for general mixtures. *Bayesian Anal.* 5:709–40

Carvalho CM, Polson NG, Scott JG. 2010. The horseshoe estimator for sparse signals. *Biometrika* 97:465–80

Chib S, Nardari F, Shephard N. 2002. Markov chain Monte Carlo methods for stochastic volatility models. *J. Econom.* 108:281–316

Conley TG, Hansen CB, McCulloch RE, Rossi PE. 2008. A semi-parametric Bayesian approach to the instrumental variable problem. *J. Econom.* 144:276–305

Conti G, Heckman JJ, Lopes HF, Piatek R. 2011. Constructing economically justified aggregates: an application of the early origins of health. *J. Econom.* In press

Dalal SR, Hall WJ. 1983. Approximating priors by mixtures of natural conjugate priors. *J. R. Stat. Soc. B* 45:278–86

Dellaportas P, Papageorgiou I. 2006. Multivariate mixtures of normals with unknown number of components. *Stat. Comput.* 16:57–68

Dey D, Müller P, Sinha D. 1998. *Practical Nonparametric and Semi-Parametric Bayesian Statistics*. New York: Springer-Verlag

DiCiccio TJ, Kass RE, Raftery A, Wasserman L. 1997. Computing Bayes factors by combining simulation and asymptotic approximations. *J. Am. Stat. Assoc.* 92:903–15

Diebolt J, Robert CP. 1994. Estimation of finite mixture distributions through Bayesian sampling. *J. R. Stat. Soc. B* 56:363–75

Escobar M, West M. 1995. Bayesian density estimation and inference using mixtures. *J. Am. Stat. Assoc.* 90:577–88

Ferguson TS. 1974. Prior distributions on spaces of probability measures. *Ann. Stat.* 2:209–30
Freedman D. 1999. On the Bernstein–von Mises theorem with infinite dimensional parameters. *Ann. Stat.* 27:1119–40
Frühwirth-Schnatter S. 2006. *Finite Mixture and Markov Switching Models*. New York: Springer
Frühwirth-Schnatter S, Lopes HF. 2010. Parsimonious Bayesian factor analysis when the number of factors is unknown. Tech. Rep., Booth School Bus., Univ. Chicago
Gamerman D, Lopes HF. 2006. *Markov Chain Monte Carlo: Stochastic Simulation for Bayesian Inference*. Boca Raton, FL: CRC. 344 pp.
Gelman A, Carlin JB, Stern HS, Rubin DB. 2004. *Bayesian Data Analysis*. Boca Raton, FL: CRC. 668 pp.
Geweke J. 1989. Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* 57:1317–39
Geweke J. 1993. Bayesian treatment of the independent student-*t* linear model. *J. Appl. Econ.* 8: S19–40
Geweke J. 1996. Bayesian reduced rank regression in econometrics. *J. Econom.* 75:121–46
Geweke J. 2005. *Comtemporary Bayesian Econometrics and Statistics*. Hoboken, NJ: Wiley. 300 pp.
Geweke J. 2007. Interpretation and inference in mixture models: Simple MCMC works. *Comput. Stat. Data Anal.* 51:3529–50
Ghosh JK, Ramamoorthi RV. 2003. *Bayesian Nonparametrics*. New York: Springer-Verlag
Ghysels E, Harvey AC, Renault E. 1996. Stochastic volatility. In *Handbook of Statistics: Statistical Methods in Finance*, ed. CR Rao, GS Maddala, pp. 119–91. Amsterdam: North-Holland
Greenberg E. 2008. *Introduction to Bayesian Econometrics*. Cambridge, UK: Cambridge Univ. Press. 205 pp.
Hjort NL, Holmes C, Müller P, Walker SG. 2010. *Bayesian Nonparametrics*. Cambridge, UK: Cambridge Univ. Press
Hull J, White A. 1987. The pricing of options on assets with stochastic volatilities. *J. Financ.* 42:281–300
Insua DR, Ruggeri F, eds. 2000. *Robust Bayesian Analysis*. New York: Springer. 422 pp.
Ishwaran H, James L. 2001. Gibbs sampling methods for stick-breaking priors. *J. Am. Stat. Assoc.* 96:161–73
Jacquier E, Polson NG, Rossi PE. 2004. Bayesian analysis of stochastic volatility models with fat-tails and correlated errors. *J. Econom.* 122:185–212
Jasra A, Holmes CC, Stephens DA. 2005. Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat. Sci.* 20:50–67
Johannes M, Polson NG. 2010. MCMC methods for continuous-time financial econometrics. In *Handbook of Financial Econometrics*, Vol. 2, ed. Y Aït-Sahalia, LP Hansen, pp. 1–72. Princeton, NJ: Princeton Univ. Press
Kalayloglu ZI, Ghosh SK. 2009. Bayesian unit-root tests for stochastic volatility models. *Stat. Methodol.* 6:189–201
Kass RE, Raftery AE. 1995. Bayes factors. *J. Am. Stat. Assoc.* 90:773–95
Kass RE, Wasserman L. 1995. A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *J. Am. Stat. Assoc.* 90:928–34
Kleibergen F, Zivot E. 2003. Bayesian and classical approaches to instrumental variable regression. *J. Econom.* 114:29–72
Kloek T, van Dijk HK. 1978. Bayesian estimates of equation system parameters: an application of integration by Monte Carlo. *Econometrica* 46:1–19
Koop G. 2003. *Bayesian Econometrics*. Chichester, UK: Wiley & Sons. 359 pp.
Koop G, Poirier DJ, Tobias JL. 2007. *Bayesian Econometric Methods*. Cambridge, UK: Cambridge Univ. Press. 357 pp.
Lancaster T. 2004. *An Introduction to Modern Bayesian Econometrics*. Malden, MA: Blackwell. 401 pp.
Liang F, Paulo R, Molina G, Clyde MA, Berger JO. 2008. Mixtures of *g* priors for Bayesian variable selection. *J. Am. Stat. Assoc.* 103:410–23
Lopes HF. 2003. Factor models. ISBA Bull. 10:7–10

Lopes HF, Carvalho CM, Johannes MJ, Polson NG. 2010. Particle learning for sequential Bayesian computation. In *Bayesian Statistics 9*, ed. JM Bernardo, MJ Bayarri, JO Berger, AP Dawid, D Heckerman, et al. New York: Oxford Univ. Press. In press

Lopes HF, Müller P, Rosner G. 2003. Bayesian meta-analysis for longitudinal data models using multivariate mixture priors. *Biometrics* 59:66–75

Lopes HF, Polson NG. 2010a. Bayesian inference for stochastic volatility modeling. In *Rethinking Risk Measurement and Reporting: Uncertainty, Bayesian Analysis and Expert Judgement*, ed. K Bocker, pp. 515–51. London: Risk Books

Lopes HF, Polson NG. 2010b. Extracting SP500 and NASDAQ volatility: the credit crisis of 2007–2008. In *The Oxford Handbook of Applied Bayesian Analysis*, ed. A O'Hagan, M West, pp. 319–42. New York: Oxford Univ. Press

Lopes HF, Polson NG. 2010c. Particle learning for fat-tailed distributions. Tech. Rep., Booth School Bus., Univ. Chicago

Lopes HF, Tsay RS. 2010. Bayesian analysis of financial time series via particle filters. *J. Forecast.* 30:168–209

Lopes HF, West M. 2004. Bayesian model assessment in factor analysis. *Stat. Sinica* 14:41–67

Marin J-M, Robert CP. 2010. On resolving the Savage-Dickey paradox. *Electron. J. Stat.* 4:643–54

McLachlan GJ, Peel D. 2000. *Finite Mixture Models*. Chichester, UK: Wiley & Sons

Müller P, Quintana FA. 2004. Nonparametric Bayesian data analysis. *Stat. Sci.* 19:95–110

Nakajima J, Omori Y. 2009. Leverage, heavy-tails and correlated jumps in stochastic volatility models. *Comput. Stat. Data Anal.* 53:2335–53

Neal R. 2000. Markov chain sampling methods for Dirichlet process mixture models. *J. Comput. Graph. Stat.* 9:249–65

Omori Y, Chib S, Shephard N, Nakajima J. 2007. Stochastic volatility with leverage: fast and efficient likelihood inference. *J. Econom.* 140:425–49

Perman M, Pitman J, Yor M. 1992. Size-biased sampling of Poisson point processes and excursions. *Probab. Theory Relat. Fields* 92:21–39

Poirier DJ. 1995. *Intermediate Statistics and Econometrics: A Comparative Approach*. Cambridge, MA: MIT Press. 715 pp.

Richardson S, Green P. 1997. On Bayesian analysis of mixtures with an unknown number of components. *J. R. Stat. Soc. B* 59:731–92

Rosenberg B. 1972. *The behaviour of random variables with nonstationary variance and the distribution of security prices*. Work. Pap. 11, Res. Program Finance, Univ. Calif., Berkeley; **http://www.haas.berkeley.edu/groups/finance/WP/rpf011.pdf**

Rossi PE, Allenby GM, McCulloch RE. 2005. *Bayesian Statistics and Marketing*. Chichester, UK: Wiley

Sims C. 2007. *Thinking about instrumental variables*. Work. Pap., Dep. Econ., Princeton Univ.; **http://www.princeton.edu/~sims/#iv**

Steel MFJ. 1998. Bayesian analysis of stochastic volatility models with flexible tails. *Econom. Rev.* 17:109–43

Stephens M. 2000. Dealing with label-switching in mixture models. *J. R. Stat. Soc. B* 62:795–809

Taylor SJ. 1986. *Modelling Financial Time Series*. New York: Wiley

Titterington DM, Smith AFM, Makov UE. 1985. *Statistical Analysis of Finite Mixture Distributions*. New York: Wiley

Tukey JW. 1978. Discussion of Granger on seasonality. In *Seasonal Analysis of Economic Time Series*, ed. A Zellner, pp. 50–53. Washington, DC: U.S. Gov. Print. Off.

Verdinelli I, Wasserman L. 1995. Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *J. Am. Stat. Assoc.* 90:614–18

Walker SG, Damien P, Laud PW, Smith AFM. 1999. Bayesian nonparametric inference for random distributions and related functions. *J. R. Stat. Soc. B* 61:485–527

West M, Harrison J. 1997. *Bayesian Forecasting and Dynamic Models*. New York: Springer. 2nd ed.