IRP Lectures **Madison, WI, August 2008**

Lecture 13, Wednesday, Aug 6th, 8.00-9.15am

Bayesian Inference

1. Introduction

In this lecture we look at Bayesian inference. Although in the statistics literature explicitly Bayesian papers take up a large proportion of journal pages these days, Bayesian methods have had very little impact in economics. This seems to be largely for historial reasons. In many empirical settings in economics Bayesian methods appear statistically more appropriate, and computationally more attractive, than the classical or frequentist methods typically used. Recent textbooks discussing modern Bayesian methods with an applied focus include Lancaster (2004) and Gelman, Carlin, Stern and Rubin (2004).

One important consideration is that in practice frequentist and Baeysian inferences are often very similar. In a regular parametric model, conventional confidence intervals around maximum likelihood (that is, the maximum likelihood estimate plus or minus 1.96 times the estimated standard error), which formally have the property that whatever the true value of the parameter is, with probability 0.95 the confidence interval covers the true value, can in fact also be interpreted as approximate Bayesian probability intervals (that is, conditional on the data and given a wide range of prior distributions, the posterior probability that the parameter lies in the confidence interval is approximately 0.95). The formal statement of this remarkable result is known as the Bernstein-Von Mises theorem. This result does not always apply in irregular cases, such as time series settings with unit roots. In those cases there are more fundamental differences between Bayesian and frequentist methods.

Typically a number of reasons are given for the lack of Bayesian methods in econometrics. One is the difficulty in choosing prior distributions. A second reason is the need for a fully specified parametric model. A third is the computational complexity of deriving posterior distributions. None of these three are compelling.

Consider first the specification of the prior distribution. In regular cases the influence of the prior distribution disappears as the sample gets large, as formalized in the Bernstein-Von Mises theorem. This is comparable to the way in which in large samples normal approximations can be used for the finite sample distributions of classical estimators. If, on the other hand, the posterior distribution is quite sensitive to the choice of prior distribution, then it is likely that the sampling distribution of the maximum likelihood estimator is not well approximated by a normal distribution centered at the true value of the parameter in a frequentist analysis.

A conventional Bayesian analysis does require a fully specified parameter model, as well as a prior distribution on all the parameters of this model. In frequentist analyses it is often possible to specify only part of the model, and use a semi-parametric approach. This advantage is not as clear cut as it may seem. When the ultimate questions of interest do not depend on certain features of the distribution, the results of a parametric model are often robust given a flexible specification of the nuisance functions. As a result, extending a semi-parametric model to a fully parametric one by flexibly modelling the nonparametric component often works well in practice. In addition, Bayesian semi-parametric methods have been developed.

Finally, traditionally computational difficulties held back applications of Bayesian methods. Modern computational advances, in particular the development of markov chain monte carlo methods, have reduced, and in many cases eliminated, these difficulties. Bayesian analyses are now feasible in many settings where they were not twenty years ago. There are now few restrictions on the type of prior distributions that can be considered and the dimension of the models used.

Bayesian methods are especially attractive in settings with many parameters. Examples discussed in these notes include panel data with individual-level heterogeneity in multiple parameters, instrumental variables with many instruments, and discrete choice data with multiple unobserved product characteristics. In such settings, methods that attempt to estimate every parameter precisely without linking it to similar parameters, often have poor repeated sampling properties. This shows up in Bayesian analyses in the dogmatic posterior distributions resulting from flat prior distributions. A more attractive approach, that is succesfuly applied in the aforementioned examples, can be based on hierarchical prior distributions where the parameters are assumed to be drawn independently from a common distribution with unknown location and scale. The recent computational advances make such models feasible in many settings.

2. Bayesian Inference

The formal set up is the following: we have a random variable X, which is known to have a probability density, or probability mass, function conditional on an unknown parameter θ. We are interested the value of the parameter θ, given one or more independent draws from the conditional distribution of X given θ. In addition we have prior beliefs about the value of the parameter θ. We will capture those prior beliefs in a prior probability distribution. We then combine this prior distribution and the sample information, using Bayes' theorem, to obtain the conditional distribution of the parameter given the data.

2.1 The General Case

Now let us do this more formally. There are two ingredients to a Bayesian analysis. First a model for the data given some unknown parameters, specified as a probability (density) function:

$$f_{X|\theta}(x|\theta).$$

As a function of the parameter this is called the likelihood function, and denoted by $\mathcal{L}(\theta)$ or $\mathcal{L}(\theta|x)$. Second, a prior distribution for the parameters, $p(\theta)$. This prior distribution is known to, that is, choosen by the researcher. Then, using Bayes' theorem we calculate the conditional distribution of the parameters given the data, also known as the posterior distribution,

$$p(\theta|x) = \frac{f_{X,\theta}(x,\theta)}{f_X(x)} = \frac{f_{X|\theta}(x|\theta)\cdot p(\theta)}{\int f_{X|\theta}(x|\theta)\cdot p(\theta)d\theta}.$$

In this step we often use a shortcut. First note that, as a function of θ, the conditional density of θ given X is proportional to

$$p(\theta|x) \propto f_{X|\theta}(x|\theta) \cdot p(\theta) = \mathcal{L}(\theta|x) \cdot p(\theta).$$

Once we calculate this product, all we have to do is found the constant that makes this expression integrate out to one as a function of the parameter. At that stage it is often easy to recognize the distribution and figure out through that route what the constant is.

2.2 A Normal Example with Unknown Mean and Known Variance, and a Single Observation

Let us look at a simple example. Suppose the conditional distribution of X given the parameter μ is normal with mean μ and variance 1, denoted by $\mathcal{N}(\mu, 1)$. Suppose we choose the prior distribution for μ to be normal with mean zero and variance 100, $\mathcal{N}(0, 100)$. What is the posterior distribution of μ given $X = x$? The posterior distribution is proportional to

$$\begin{aligned} f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2\cdot 100}\mu^2\right) \\ &= \exp -\frac{1}{2}\Big(x^2 - 2x\mu + \mu^2 + \mu^2/100\Big) \\ &\propto \exp\left(-\frac{1}{2(100/101)}\Big(\mu - (100/101)x\Big)^2\right). \end{aligned}$$

This implies that the conditional distribution of μ given $X = x$ is normal with mean $(100/101) \cdot x$ and variance $100/101$, or $\mathcal{N}(x \cdot 100/101, 100/101)$.

In this example the model was a normal distribution for X given the unknown mean μ, and we choose a normal prior distribution. This was a very convenient choice, leading the posterior distribution to be normal as well. In this case the normal prior distribution is a conjugate prior distribution, implying that the posterior distribution is in the same family of distributions as the prior distribution, allowing for analytic calculations. If we had choosen a different prior distribution it would typically not have been possible to obtain an analytic expression for the posterior distribution.

Let us continue the normal distribution example, but generalize the prior distribution. Suppose that given μ the random variable X has a normal distribution with mean μ and known variance σ^2, or $\mathcal{N}(\mu, \sigma^2)$. The prior distribution for μ is normal with mean μ_0 and variance τ^2, or $\mathcal{N}(\mu_0, \tau^2)$. Then the posterior distribution is proportional to:

$$\begin{aligned}
f_{\mu|X}(\mu|x) &\propto \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \cdot \exp\left(-\frac{1}{2\cdot\tau^2}(\mu-\mu_0)^2\right) \\
&\propto \exp -\frac{1}{2}\left(\frac{x^2}{\sigma^2} - \frac{2x\mu}{\sigma^2} + \frac{\mu^2}{\sigma^2} + \frac{\mu^2}{\tau^2} - \frac{2\mu\mu_0}{\tau^2} + \frac{\mu_0^2}{\tau^2}\right) \\
&\propto \exp -\frac{1}{2}\left(\mu^2\frac{\sigma^2+\tau^2}{\tau^2\sigma^2} - \mu\frac{2x\tau^2+2\mu_0\sigma^2}{\tau^2\cdot\sigma^2}\right) \\
&\propto \exp -\frac{1}{2(1/(1/\tau^2+1/\sigma^2))}\left((\mu - (x/\sigma^2+\mu_0/\tau^2)/(1/\sigma^2+1/\tau^2)\right) \\
&\sim \mathcal{N}\left(\frac{x/\sigma^2+\mu_0/\tau^2}{1/\sigma^2+1/\tau^2}, \frac{1}{1/\tau^2+1/\sigma^2)}\right).
\end{aligned}$$

The result is quite intuitive: the posterior mean is a weighted average of the prior mean μ_0 and the observation x with weights adding up to one and proportional to the precision (defined as one over the variance), $1/\sigma^2$ for x and $1/\tau^2$ for μ_0:

$$\mathbb{E}[\mu|X=x] = \frac{\frac{x}{\sigma^2} + \frac{\mu_0}{\tau^2}}{\frac{1}{\sigma^2} + \frac{1}{\tau^2}}.$$

The posterior precision is obtained by adding up the precision for each component

$$\frac{1}{\mathbb{V}(\mu|X)} = \frac{1}{\sigma^2} + \frac{1}{\tau^2}.$$

So, what you expect *ex post*, $\mathbb{E}[\mu|X]$, that is, after seeing the data, is a weighted average of what you expected before seeing the data, $\mathbb{E}[\mu] = \mu_0$, and the observation, X, with the weights determined by their respective variances.

There are a number of insights obtained by studying this example more carefully. Suppose we are really sure about the value of μ before we conduct the experiment. In that case we would set τ^2 small and the weight given to the observation would be small, and the posterior distribution would be close to the prior distribution. Suppose on the other hand we are very unsure about the value of μ. What value for τ should we choose? Obviously a large value, but what is the limit? We can in fact let τ go to infinity. Even though the prior distribution is not a proper distribution anymore if $\tau^2 = \infty$, the posterior distribution is perfectly well defined, namely $\mu|X \sim \mathcal{N}(X, \sigma^2)$. In that case we have an improper prior distribution. We give equal prior weight to any value of μ. That would seem to capture pretty well the idea that a priori we are ignorant about μ. This is the idea of looking for an relatively uninformative prior distribution. This is not always easy, and the subject of a large literature. For example, a flat prior distribution is not always uninformative about particular functions of parameters.

2.3 A Normal Example with Unknown Mean and Known Variance and Multiple Observations

Now suppose we have N independent draws from a normal distribution with unknown mean μ and known variance σ^2. Suppose we choose, as before, the prior distribution to be normal with mean μ_0 and variance τ^2, or $\mathcal{N}(\mu_0, \tau^2)$.

The likelihood function is

$$\mathcal{L}(\mu|\sigma^2, x_1, \ldots, x_N) = \prod_{i=1}^{N} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right),$$

so that with a normal (μ_0, τ^2) prior distribution the posterior distribution is proportional to

$$p(\mu|x_1, \ldots, x_N) \propto \exp\left(-\frac{1}{2\tau^2}(\mu - \mu_0)^2 \right) \cdot \prod_{i=1}^{N} \exp\left(-\frac{1}{2\sigma^2}(x_i - \mu)^2 \right).$$

Thus, with N observations $x_1, \ldots, x_N$ we find, after straightforward calculations,

$$\mu|X_1, \ldots, X_N \sim \mathcal{N}\left(\frac{\mu_0/\tau^2 + \sum x_i/\sigma^2}{1/\tau^2 + N/\sigma^2}, \frac{1}{1/\tau^2 + N/\sigma^2} \right).$$

2.4 The Normal Distribution with Known Mean and Unknown Variance

Let us also briefly look at the case of a normal model with known mean and unknown variance. Thus,

$$X_i|\sigma^2 \sim \mathcal{N}(0, \sigma^2),$$

and $X_1, \ldots, X_N$ independent given σ^2. The likelihood function is

$$\mathcal{L}(\sigma^2) = \prod_{i=1}^{N} \sigma^{-N} \exp\left(-\frac{1}{2\sigma^2} X_i^2 \right).$$

Now suppose that the prior distribution for σ^2 is, for some fixed S_0^2 and K_0, such that the distribution of $\sigma^{-2} \cdot S_0^2 \cdot K_0$ is chi-squared with K_0 degrees of freedom. In other words, the prior distribution of σ^{-2} is $(S_0^2 \cdot K_0)^{-1}$ times a chi-squared distribution with K_0 degrees of freedom. Then the posterior distribution of σ^{-2} is $(S_0^2 \cdot K_0 + \sum_i X_i^2)^{-1}$ times a chi-squared distribution with $K_0 + N$ degrees of freedom, so this is a conjugate prior distribution.

3. The Bernstein-Von Mises Theorem

Let us go back to the normal example with N observations, and unknown mean and known variance. In that case with a normal $\mathcal{N}(\mu_0, \tau^2)$ prior distribution the posterior for μ is

$$\mu|x_1, \ldots, x_N \sim \mathcal{N}\left(\overline{x} \cdot \frac{1}{1 + \sigma^2/(N \cdot \tau^2)} + \mu_0 \cdot \frac{\sigma^2/(N\tau^2)}{1 + \sigma^2/(N\tau^2)}, \frac{\sigma^2/N}{1 + \sigma^2/(N\tau^2)}\right).$$

When N is very large, the distribution of $\sqrt{N}(\mu - \bar{x})$ conditional on the data is approximately

$$\sqrt{N}(\bar{x} - \mu)|x_1, \ldots, x_N \sim \mathcal{N}(0, \sigma^2).$$

In other words, in large samples the influence of the prior distribution disappears, unless the prior distribution is choosen particularly badly, e.g., τ^2 equal to zero. This is true in general, i.e., for different models and different prior distributions. Moreover, in a frequentist analysis we would find that in large samples (and in this specific normal example even in finite samples),

$$\sqrt{N}(\bar{x} - \mu)|\mu \sim \mathcal{N}(0, \sigma^2).$$

Let us return to the Bernoulli example to see the same point. Suppose that conditional on the parameter $P = p$, the random variables $X_1, X_2, \ldots, X_N$ are independent with Bernoulli distributions with probability p. Let the prior distribution of P be Beta with parameters α and β, or $\mathcal{B}(\alpha, \beta)$. Now consider the conditional distribution of P given $X_1, \ldots, X_N$:

$$f_{P|X_1,\ldots,X_N}(p|x) \propto p^{\alpha - 1 + \sum_{i=1}^N X_i} \cdot (1 - p)^{\beta - 1 + N - \sum_{i=1}^N X_i},$$

which is a Beta distribution, $\mathcal{B}(\alpha - 1 + \sum_{i=1}^N X_i, \beta - 1 + N - \sum_{i=1}^N X_i)$, with mean and variance

$$\mathbb{E}[P|X_1,\ldots,X_N] = \frac{\alpha + \sum_{i=1}^N X_i}{\alpha+\beta+N}, \qquad \text{and } \mathbb{V}(P) = \frac{(\alpha + \sum_{i=1}^N X_i)(\beta + N - \sum_{i=1}^N X_i)}{(\alpha+\beta+N)^2(\alpha+\beta+1+N)}.$$

What happens if N gets large? Let $\hat{p} = \sum_i X_i/N$ be the relative frequency of success (which is the maximum likelihood estimator for p). Then the mean and variance converge to

$$\mathbb{E}[P|X_1,\ldots,X_N] \approx \hat{p},$$

and

$$\mathbb{V}(P) \approx 0.$$

As the sample size gets larger, the posterior distribution becomes concentrated at a value that does not depend on the prior distribution. This in fact can be taken a step further. In this example, the limiting distribution of $\sqrt{N}\cdot(P-\hat{p})$ conditional on the data, can be shown to be

$$\sqrt{N}(\hat{p} - P|x_1,\ldots,x_N \xrightarrow{d} \mathcal{N}(0,\hat{p}(1-\hat{p})),$$

again irrespective of the choice of α and β. The interpretation of this result is very important: in large sample the choice of prior distribution is not important in the sense that the information in the prior distribution gets dominated by the sample information. That is, unless your prior beliefs are so strong that they cannot be overturned by evidence (i.e., the prior distribution is zero over some important range of the parameter space), at some point the evidence in the data outweights any prior beliefs you might have started out with.

This is known as the Bernstein-von Mises Theorem. Here is a general statement for the scalar case. Let the information matrix $\mathcal{I}_\theta$ at θ:

$$\mathcal{I}_\theta = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta\partial\theta'}\ln f_X(x|\theta)\right] = -\int \frac{\partial^2}{\partial\theta\partial\theta'}\ln f_X(x|\theta) f_X(x|\theta)dx,$$

and let σ^2 be the inverse at a fixed value θ_0.

$$\sigma^2 = \mathcal{I}_{\theta_0}^{-1}.$$

Let $p(\theta)$ be the prior distribution, and $p_{\theta|X_1,\ldots,X_N}(\theta|X_1,\ldots,X_N)$ be the posterior distribution. Now let us look at the distribution of a transformation of θ, $\gamma = \sqrt{N}(\theta - \theta_0)$, with density $p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) = p_{\theta|X_1,\ldots,X_N}(\theta_0 + \sqrt{N}\cdot\gamma|X_1,\ldots,X_N)/\sqrt{N}$. Now let us look at the posterior distribution for θ if in fact the data were generated by $f(x|\theta)$ with $\theta = \theta_0$. In that case the posterior distribution of γ converges to a normal distribution with mean zero and variance equal to σ^2 in the sense that

$$\sup_{\gamma}\left|p_{\gamma|X_1,\ldots,X_N}(\gamma|X_1,\ldots,X_N) - \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{1}{2\sigma^2}\gamma^2\right)\right| \longrightarrow 0.$$

See Van der Vaart (2001), or Ferguson (1996). At the same time, if the true value is θ_0, then the mle $\hat{\theta}_{mle}$ also has a limiting distribution with mean zero and variance σ^2:

$$\sqrt{N}(\hat{\theta}_{ml} - \theta_0) \xrightarrow{d} \mathcal{N}(0, \sigma^2).$$

The implication is that we can interpret confidence intervals as approximate probability intervals from a Bayesian perspective. Specifically, let the 95% confidence interval be $[\hat{\theta}_{ml} - 1.96\cdot\hat{\sigma}/\sqrt{N}, \hat{\theta}_{ml} + 1.96\cdot\hat{\sigma}/\sqrt{N}]$. Then, approximately,

$$\Pr\left(\hat{\theta}_{ml} - 1.96\cdot\hat{\sigma}/\sqrt{N} \leq \theta \leq \hat{\theta}_{ml} + 1.96\cdot\hat{\sigma}/\sqrt{N}\middle| X_1,\ldots,X_N\right) \longrightarrow 0.95.$$

There are important cases where this result does not hold, typically when convergence to the limit distribution is not uniform. One is the unit-root setting. In a simple first order autoregressive example it is still the case that with a normal prior distribution for the autoregressive parameter the posterior distribution is normal (see Sims and Uhlig, 1991). However, if the true value of the autoregressive parameter is unity, the sampling distribution is not normal even in large samples. In that case one has to take a more principled stand whether one wants to make subjective probability statements, or frequentist claims.

4. Markov Chain Monte Carlo Methods

Are we really restricted to choosing the prior distributions in these conjugate families as we did in the examples so far? No. The posterior distributions are well defined irrespective of conjugacy. Conjugacy only simplifies the computations. If you are outside the conjugate families, you typically have to resort to numerical methods for calculating posterior moments. Recently many methods have been developed that make this process much easier, including *Gibbs* sampling, *Data Augmentation*, and the *Metropolis-Hastings* algorithm. All three are examples of *M*arkov-Chain-Monte-Carlo or MCMC methods.

The general idea is to construct a chain, or sequence of values, $\theta_0, \theta_1, \ldots$, such that for large k θ_k can be viewed as a draw from the posterior distribution of θ given the data. This is implemented through an algorithm that, given a current value of the parameter vector θ_k, and given the data $X_1, \ldots, X_N$ draws a new value θ_{k+1} from a distribution $f(\cdot)$ indexed by θ_k and the data:

$$\theta_{k+1} \sim f(\theta|\theta_k, \text{data}),$$

in such a way that if the original θ_k came from the posterior distribution, then so does θ_{k+1} (although θ_k and θ_{k+1} in general will not be independent draws)

$$\theta_k|\text{data} \sim p(\theta|\text{data}), \qquad \text{then} \;\; \theta_{k+1}|\text{data} \sim p(\theta|\text{data}).$$

Even if we have such an algorithm, the problem is that in principle we would need a starting value θ_0 that such that

$$\theta_0 \sim p(\theta|\text{data}).$$

However, in many cases, irrespective of where we start, that is, irrespective of θ_0, as $k \longrightarrow \infty$, it will be the case that the distribution of the parameter conditional only on the data converges to the posterior distribution:

$$\theta_k|\text{data} \xrightarrow{d} p(\theta|\text{data}),$$

as $k \longrightarrow \infty$.

If that is true, then we can just pick a θ_0, run the chain for a long time, collect the values $\theta_0, \ldots, \theta_K$ for a large value of K, and approximate the posterior distribution by the distribution of $\theta_{K_0}, \ldots, \theta_K$. For example, the mean and standard deviation of the posterior distribution would be estimated as

$$\hat{\mathbb{E}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \theta_k,$$

and

$$\hat{\mathbb{V}}[\theta|\text{data}] = \frac{1}{K - K_0 + 1} \sum_{k=K_0}^{K} \left(\theta_k - \hat{\mathbb{E}}[\theta|\text{data}]\right)^2.$$

The first $K_0 - 1$ iterations are discarded to let algorithm converge to the stationary distribution, or "burn in."

4.1 Gibbs Sampling

The idea being the Gibbs sampler is to partition the vector of parameters θ into two (or more) parts, $\theta' = (\theta_1', \theta_2')$. Instead of sampling θ_{k+1} directly from a conditional distribution of

$$f(\theta|\theta_k, \text{data}),$$

we first sample $\theta_{1,k+1}$ from the conditional distribution of

$$p(\theta_1|\theta_{2,k}, \text{data}),$$

and then sample $\theta_{2,k+1}$ from the conditional distribution of

$$p(\theta_2|\theta_{1,k+1}, \text{data}).$$

It is clear that if $(\theta_{1,k}, \theta_{2,k})$ is from the posterior distribution, then so is $(\theta_{1,k}, \theta_{2,k})$.

4.2 Data Augmentation

This is best illustrated with an example. Suppose we are interested in estimating the parameters of a censored regression or Tobit model. There is a latent variable

$$Y_i^* = X_i'\beta + \varepsilon_i,$$

with $\varepsilon_i|X_i \sim \mathcal{N}(0, 1)$. (I assume the variance is known for simplicity. This is not essential). We observe

$$Y_i = \max(0, Y_i^*),$$

and the regressors X_i. Suppose the prior distribution for β is normal with some mean μ, and some covariance matrix Ω.

The posterior distribution for β does not have a closed form expression. This is not due to an awkward choice for the prior distribution, because there is no conjugate family for this problem. There is however a simple way of obtaining draws from the posterior distribution using data augmentation in combination with the Gibbs sampler. The first key insight is to view both the vector $\mathbf{Y}^* = (Y_1^*, \ldots, Y_N^*)$ and β as unknown random variables. The Gibbs sampler consists of two steps. First we draw all the missing elements of $\mathbf{Y}^*$ given the current value of the parameter β, say β_k. This involves drawing a series of truncated univariate normal random variables:

$$Y_i^*|\beta, \text{data} \sim \mathcal{TN}\left(X_i'\beta, 1, 0\right),$$

if observation i is truncated, where $\mathcal{TN}(\mu, \sigma^2, c)$ denotes a truncated normal distribution with mean and variance (for the not truncated normal distribution) μ and σ^2, and truncation point c (truncated from above). (Note that we do not need to draw the observed values of Y_i^*.) Second, we draw a new value for the parameter, β_{k+1} given the data and given the (partly drawn) $\mathbf{Y}^*$. The latter is easy given the normal prior distribution: the posterior is normal:

$$p\left(\beta|\text{data}, \mathbf{Y}^*\right) \sim \mathcal{N}\left(\left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1} \cdot \left(\mathbf{X}'\mathbf{Y} + \Omega^{-1}\mu\right), \left(\mathbf{X}'\mathbf{X} + \Omega^{-1}\right)^{-1}\right).$$

In this example it would still have been feasible to do evaluate the likelihood function directly using numerical integration. Another example where the computational advantages of using data augmentation are even more striking is the multinomial probit model with an unrestricted covariance matrix. See Rossi, Allenby and McCulloch (2005).

4.3 Metropolis-Hastings

We are again interested in $p(\theta|\text{data})$. In this case $p(\theta|\text{data})$ is assumed to be easy to evaluate, but difficult to draw from. Suppose we have a current value θ_k. Then we draw a new candidate value for the chain from a candidate distribution $q(\theta|\theta_k, \text{data})$. This distribution may (but need not) depend on θ_k. Denote the candidate value by θ. We will either accept the new value, or keep the old value. Then we calculate the ratio

$$r(\theta_k, \theta) = \frac{p(\theta|\text{data}) \cdot q(\theta_k|\theta, \text{data})}{p(\theta_k|\text{data}) \cdot q(\theta|\theta_k, \text{data})}.$$

The probability that the new draw θ is accepted is

$$\rho(\theta_k, \theta) = \min\left(1, r(\theta_k, \theta)\right),$$

so that

$$\Pr\left(\theta_{k+1} = \theta\right) = \rho(\theta_k, \theta), \qquad \text{and } \Pr\left(\theta_{k+1} = \theta_k\right) = 1 - \rho(\theta_k, \theta).$$

The optimal choice for the candidate distribution is

$$q^*(\theta|\theta_k, \text{data}) = p(\theta|\text{data}),$$

so that $\rho(\theta_k, \theta) = 1$ and every draw will get accepted. The trouble is that it is difficult to draw from this distribution. In practice you want to have a relatively dispersed distribution as the candidate distribution, so that the ratio $r(\theta_k, \theta)$ does not get too large.

5. Examples

Here we discuss a number of applications of Bayesian methods. All models contain parameters that are difficult to estimate consistently, and in all cases numerical methods are required to obtain draws from the posterior distribution. The first two are about random coeffiecients. In that case Bernstein-Von Mises would only apply to the individual level parameters if the number of observations per individual would get large.

5.1 Demand Models with Unobserved Heterogeneity in Preferences

Rossi, McCulloch, and Allenby (1996, RMA) are interested in the optimal design of coupon policies. Supermarkets can choose to offer identical coupons for a particular product (tuna cans is the example they use). Alternatively, they may choose to offer differential coupons based on consumer's fixed characteristics. Taking this ever further, they could make the value of the coupon a function of the purchase history of the individual, for example tailoring the amount of the discount offered in the coupon to the evidence for price sensitivity contained in purchase patterns. RMA set out to estimate the returns to various coupon policies. It is clear that for this investigation to be meaningful one needs to allow for household-level heterogeneity in taste parameters and price elasticities. Even with large amounts of data available, there will be many households for whom these parameters cannot be estimated precisely. RMA therefore use a hieararchical or random coefficients model.

RMA model households choosing the product with the highest utility, where utility for household i, product j, $j = 0, 1, \ldots, J$, at purchase time t is

$$U_{ijt} = X_{it}'\beta_i + \epsilon_{ijt},$$

with the ϵ_{ijt} independent accross households, products and purchase times, and normally distributed with product-specific variances σ_j^2 (and σ_0^2 normalized to one). The X_{it} are observed choice characteristics that in the RMA application include price, some marketing variables, as well as brand dummies. All choice characteristics are assumed to be exogenous, although that assumption may be questioned for the price and marketing variables. Because for some households we have few purchases, it is not possible to accurately estimate all β_i parameters. RMA therefore assume that the household-specific taste parameters are random draws from a normal distribution centered at $Z_i'\Gamma$:

$$\beta_i = Z_i'\Gamma + \eta_i, \qquad \eta_i \sim \mathcal{N}(0, \Sigma).$$

Now Gibbs sampling can be used to obtain draws from the posterior distribution of the β_i. To be a little more precise, let us describe the steps in the Gibbs sampler for this example. For more details see RMA. RMA use a normal prior distribution for Γ, a Wishart prior distribution for Σ^{-1}, and inverse Gamma prior distributions for the σ_j^2. To implement the Gibbs sampler, the key is to treat the unobserved utilities as parameters.

The first step is to draw the household parameters β_i given the utilities U_{ijt} and the common parameters Γ, Σ, and σ_j^2. This is straightforward, because we have a standard normal linear model for the utilities, with a normal prior distribution for β_i with parameters

$Z_i'\Gamma$ and variance Σ, and T_i observations. We can draw from this posterior distribution for each household i.

In the second step we draw the σ_j^2 using the results for the normal distribution with known mean and unknown variance.

The third step is to draw from the posterior of Γ and Σ, given the β_i. This again is just a normal linear model, now with unknown mean and unknown variance.

The fourth step is to draw the unobserved utilities given the β_i and the data. Doing this one household/choice at a time, conditioning on the utilities for the other choices, this merely involves drawing from a truncated normal distribution, which is simple and fast.

For some households, those with many recorded purchases and sufficient variation in product characteristics, the posterior distribution will be tight, whereas for others there may be little information in the data and the posterior distribution, conditional on the data as well as Γ and Σ, will essentially be the prior distribution for β_i, which is $\mathcal{N}(Z_i'\Gamma, \Sigma)$.

To think about optimal coupon policies given a particular information set it is useful to think first about the posterior distribution of the household specific parameters β_i. If a supermarket had full information about the household parameters β_i, there would be no additional value in the household characteristics or the purchase history. When we therefore compare a blanket coupon policy (where every household would receive a coupon with the same value) with one that depends on a larger information set that household demographics, or one that also includes purchase histories, the key question is how much precision the information adds about the household level parameters. Specifically, how does the marginal distribution of the household parameters compare with the conditional distribution given purchase histories or given demographics. To make this specific, suppose that the there is only one choice characteristic, price, with household parameter β_i.

The starting point is the case with no household information whatsoever. We can simulate draws from this distribution by drawing from the conditional distribution of β_i given the data for randomly selected households. In the second case we allow conditioning on the household demographic characteristics Z_i. This leads to less dispersed posterior distributions for the price coefficients. In the third case we also condition on purchase histories. Figure 1, taken from RMA shows for ten households the boxplots of the posterior distribution of the price coefficient under these information sets, one can see the increased precision that results from conditioning on the purchase histories.

5.2 Panel Data Models with Multiple Individual Specific Parameters

Chamberlain and Hirano (1999, CH), see also Hirano (2002), are interested in deriving predictive distributions for earnings using longitudinal data. They are particularly interested in allowing for unobserved individual-level heterogeneity in earnings variances. The specific model they use assumes that log earnings Y_{it} follow the process

$$Y_{it} = X_i'\beta + V_{it} + \alpha_i + U_{it}/h_i.$$

The key innovation in the CH study is the individual variation in the conditional variance, captured by h_i. In this specification $X_i'\beta$ is a systematic component of log earnings, similar to that in specifications used in Abowd and Card () (CH actually use a more general non-linear specification, but the simpler one suffices for the points we make here.) The second component in the model, V_{it}, is a first order autoregressive component,

$$V_{it} = \gamma \cdot V_{it-1} + W_{it},$$

where

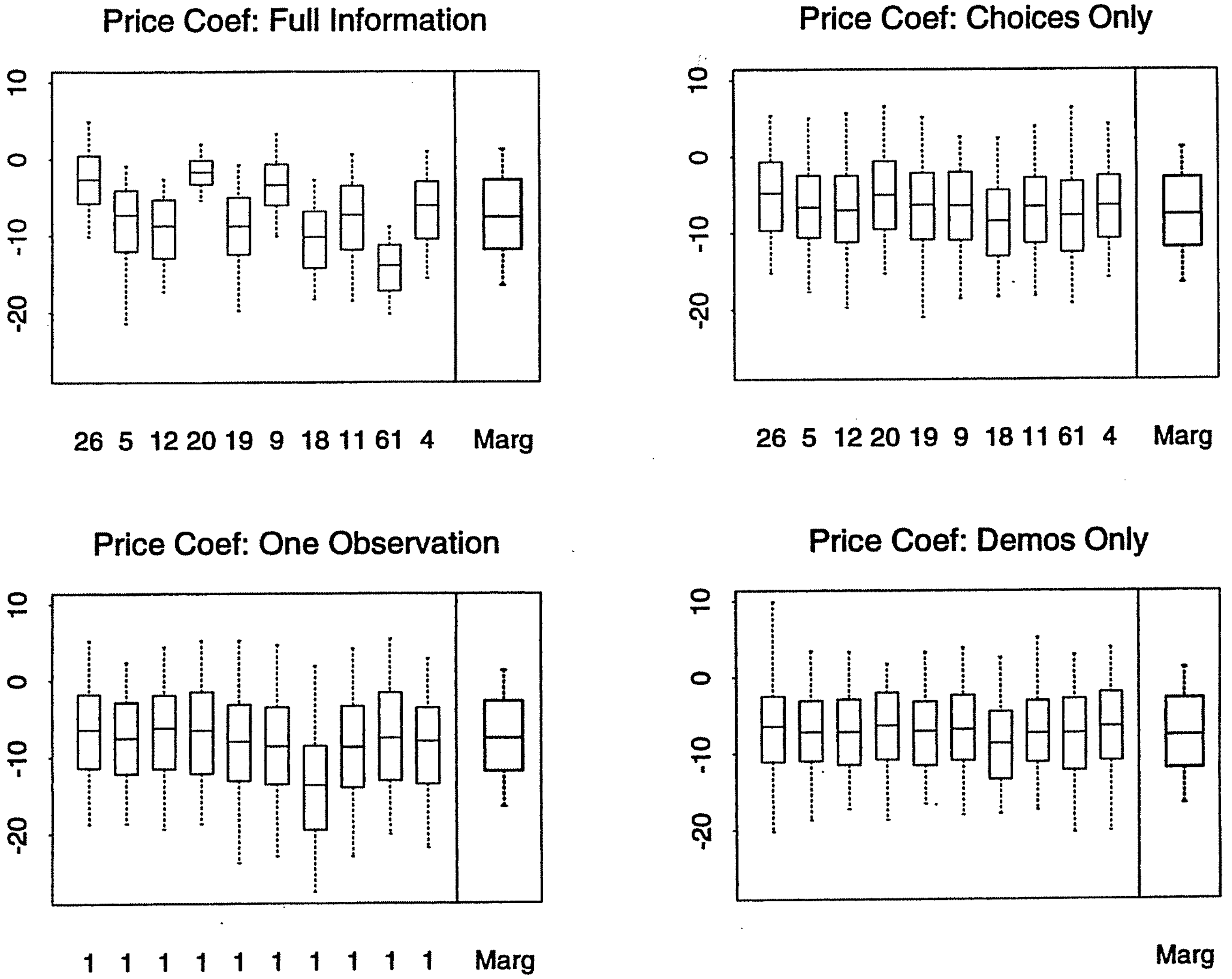
$$V_{i1} \sim \mathcal{N}(0, \sigma_v^2), \qquad W_{it} \sim \mathcal{N}(0, \sigma_w^2).$$

The first factor in the last component has a standard normal distribution,

$$U_{it} \sim \mathcal{N}(0, 1).$$

Analyzing this model by attempting to estimate the α_i and h_i directly would be misguided. From a Bayesian perspective this corresponds to assuming a flat prior distribution on a high-dimensional parameter space.

Figure 2 Boxplots of posterior distributions of household price coefficients. Various information sets. 10 selected households with the number of purchase occasions indicated along the *X* axis below each boxplot. The boxplot labelled "Marg" is the predictive distribution for a representative household from the model heterogeneity distribution. Note that these are the 11–20th households as ordered in our dataset.



To avoid such pitfalls CH model α_i and h_i through a random effects specification.

$$\alpha_i \sim \mathcal{N}(0,\sigma_\alpha^2). \qquad \text{and} \quad h_i \sim \mathcal{G}(m/2,\tau/2).$$

In their empirical application using data from the Panel Study of Income Dynamics (PSID), CH find strong evidence of heterogeneity in conditional variances. Some of this heterogeneity is systematically associated with observed characteristics of the individual such as education, with higher educated individuals experiences lower levels of volatility. Much of the heterogeneity, however, is within groups homogenous in observed characteristics.

The following table, from CH, presents quantiles of the predictive distribution of the conditional standard deviation $1/\sqrt{h_i}$ for different demographic groups: Up to here one could have done essentially the same using frequentist methods. One could estimate first the common parameters of the model, β, σ_v^2, σ_w^2, m, τ, and σ_α^2 by maximum likelihood given the specification of the model. Conditional on the covariates one could for each demographic group write the quantiles of the conditional standard deviation in terms of these parameters and obtain point estimates for them.

Table 1: QUANTILES OF THE PREDICTIVE DISTRIBUTION OF THE CONDITIONAL STANDARD DEVIATION

	Quantile						
Sample	0.05	0.10	0.25	0.50	0.75	0.90	0.95
All (N=813)	0.04	0.05	0.07	0.11	0.20	0.45	0.81
High School Dropouts (N=37)	0.06	0.08	0.11	0.16	0.27	0.49	0.79
High School Graduates (N=100)	0.04	0.05	0.06	0.11	0.21	0.49	0.93
College Graduates (N=122)	0.03	0.04	0.05	0.09	0.18	0.40	0.75

However, CH wish to go beyond this and infer individual-level predictive distributions for earnings. Taking a particular individual, one can derive the posterior distribution of α_i, h_i, β, σ_v^2, and σ_w^2, given that individual's earnings as well as other earnings, and predict future earnings. To illustrate this CH report earnings predictions for a number of individuals.

Taking two of their observations, one an individual with a sample standard deviation of log earnings of 0.07 and one an individual with a sample standard deviation of 0.47, they report the difference between the 0.90 and 0.10 quantile for the log earnings distribution for these individuals 1 and 5 years into the future.

Table 2:

		0.90-0.10 quantile	
individual	sample std	1 year out	5 years out
321	0.07	0.32	0.60
415	0.47	1.29	1.29

The variation reported in the CH results may have substantial importance for variation in optimal savings behavior by individuals.

5.3 Instrumental Variables with Many Instruments

In Chamberlain and Imbens (1995, CI) analyze the many instrument problem from a Bayesian perspective. CI use the reduced form for years of education,

$$X_i = \pi_0 + Z_i'\pi_1 + \eta_i,$$

combined with a linear specification for log earnings,

$$Y_i = \alpha + \beta \cdot Z_i'\pi_1 + \varepsilon_i.$$

CI assume joint normality for the reduced form errors,

$$\begin{pmatrix} \varepsilon_i \\ \eta_i \end{pmatrix} \sim \mathcal{N}(0, \Omega).$$

This gives a likelihood function

$$\mathcal{L}(\beta, \alpha, \pi_0, \pi_1, \Omega | \text{data}).$$

The focus of the CI paper is on inference for β, and the sensitivity of such inferences to the choice of prior distribution in settings with large numbers of instruments. In that case the dimension of the parameter space is high. Hence a flat prior distribution may in fact be a poor choice. One way to illustrate see this is that a flat prior on π_1 leads to a prior on the sum $\sum_{k=1}^{K} \pi_{ik}^2$ that puts most probability mass away from zero. If in fact the concern is that collectively, the instruments are all weak, one should allow for this possibility in the prior distribution. To be specific, if the prior distribution for the π_{1k} is dispersed, say $\mathcal{N}(0, 100^2)$, then the prior distribution for the $\sum_i \pi_{1k}^2$ is 100 times a chi-squared random variable with degrees of freedom equal to K, implying that *a priori* the concentration parameter is known to be large.

CI then show that the posterior distribution for β, under a flat prior distribution for π_1 provides an accurate approximation to the sampling distribution of the TSLS estimator, providing both a further illustration of the lack of appeal of TSLS in settings with many instruments, and the unattractiveness of the flat prior distribution.

As an alternative CI suggest a hierarchical prior distribution with

$$\pi_{1k} \sim \mathcal{N}(\mu_\pi, \sigma_\pi^2).$$

In the Angrist-Krueger 1991 compulsory schooling example there is in fact a substantive reason to believe that σ_π^2 is small. If the π_{1k} represent the effect of the differences in the amount of required schooling, one would expect the magnitude of the π_{1k} to be less than the amount of variation in the compulsory schooling. The latter is less than one year. Since any distribution with support on $[0, 1]$ has a variance less than or equal to $1/12$, the standard deviation of the first stage coefficients should not be more than $\sqrt{1/12} = 0.289$. Using the Angrist-Krueger data CI find that the posterior distribution for σ_π is concentrated close to zero, with the posterior mean and median equal to 0.119.

5.4 Binary Response with Endogenous Discrete Regressors

Geweke, Gowrisankaran, and Town (2003, GGT) are interested in estimating the effect of hospital quality on mortality, taking into account possibly non-random selection of patients into hospitals. Patients can choose from 114 hospitals. Given their observed individual characteristics Z_i, latent mortality is

$$Y_i^* = \sum_{j=1}^{113} C_{ij}\beta_j + Z_i'\gamma + \epsilon_i,$$

where C_{ij} is an indicator for patient i going to hospital j. The focus is on the hospital effects on mortality, β_j. Realized mortality is

$$Y_i = 1\{Y_i^* \geq 0\}.$$

The concern is about selection into the hospitals, and the possibility that this is related to unobserved components of latent mortality GGT model latent the latent utility for patient i associated with hospital j as

$$C_{ij}^* = X_{ij}'\alpha + \eta_{ij},$$

where the X_{ij} are hospital-individual specific characteristics, including distance to hospital. Patient i then chooses hospital j if

$$C_{ij}^* \geq C_{ik}, \quad \text{for } k = 1, \ldots, 114.$$

The endogeneity is modelled through the potential correlation between η_{ij} and ϵ_i. Specifically, GGT asssume that as

$$\epsilon_i = \sum_{j=1}^{113} \eta_{ij} \cdot \delta_j + \zeta_i,$$

where the ζ_i is a standard normal random variable, independent of the other unobserved components. GGT model the η_{ij} as standard normal, independent across hospitals and across individuals. This is a very strong assumption, implying essentially the independence of irrelevant alternatives property. One may wish to relax this by allowing for random coefficients on the hospital characteristics.

Given these modelling decisions GGT have a fully specified joint distribution of hospital choice and mortality given hospital and individual characteristics. The log likelihood function is highly nonlinear, and it is unlikely it can be well approximated by a quadratic function. GGT therefore use Bayesian methods, and in particular the Gibbs sampler to obtain draws from the posterior distribution of interest. In their empirical analysis GGT find strong evidence for non-random selection. They find that higher quality hospitals attract sicker patients, to the extent that a model based on exogenous selection would have led to misleading conclusions on hospital quality.

5.5 Discrete Choice Models with Unobserved Choice Characteristics

Athey and Imbens (2007, AI) study discrete choice models, allowing both for unobserved individual heterogeneity in taste parameters as well as for multiple unobserved choice characteristics. In such settings the likelihood function is multi-modal, and frequentist approximations based on quadratic approximations to the log likelihood function around the maximum likelihood estimator are unlikely to be accurate. The specific model AI use assumes that the utility for individual i in market t for choice j is

$$U_{ijt} = X_{it}'\beta_i + \xi_j'\gamma_i + \epsilon_{ijt},$$

where X_{it} are market-specific observed choice characteristics, ξ_j is a vector of unobserved choice characteristics, and ϵ_{ijt} is an idiosyncratic error term, independent accross market, choices, and individuals, with a normal distribution centered at zero, and with the variance normalized to unity. The individual-specific taste parameters for both the observed and unobserved choice characteristics normally distributed:

$$\begin{pmatrix} \beta_i \\ \gamma_i \end{pmatrix} |Z_i \sim \mathcal{N}(\Delta Z_i, \Omega),$$

with the Z_i observed individual characteristics.

AI specify a prior distribution on the common parameters, Δ, and Ω, and on the values of the unobserved choice characteristics ξ_j. Using gibbs sampling and data augmentation with the unobserved utilities as unobserved random variables makes sampling from the posterior distribution conceptually straightforward even in cases with more than one unobserved choice characteristic. In contrast, earlier studies using multiple unobserved choice characteristics

(Elrod and Keane, 1995; Goettler and Shachar, 2001), using frequentist methods, faced much heavier computational burdens.

References

Athey, S., and G. Imbens, (2007), "Discrete Choice Models with Multiple Unobserved Product Characteristics," *International Economic Review*, forthcoming.

Box, G., and G. Tiao, (1973), *Bayesian Inference in Statistical Analysis,* Wiley, NY.

Chamberlain, G., and K. Hirano, (1996), "Hirearchical Bayes Models with Many Instrumental Variables," NBER Technical Working Paper 204.

Chamberlain, G., and G. Imbens, (1999), "Predictive Distributions based on Longitudinal Eearnings Data," *Annales d'Economie et de Statistique*, 55-56, 211-242.

Elrod, T., and M. Keane, (1995), "A Factor-Analytic Probit Model for Representing the Market Structure in Panel Data, "*Journal of Marketing Research*, Vol. XXXII, 1-16.

Ferguson, T., (1996), *A Course in Large Sample Theory*, Chapman and Hall, new York, NY.

Gelman, A., J. Carlin, H. Stenr, and D. Rubin, (2004), *Bayesian Data Analysis,* Chapman and Hall, New York, NY.

Gelman, A., and J. Hill, (2007), *Data Analysis Using Regression and Multilevel/Hierarchical Models,* Cambridge University Press.

Geweke, J., G. Gowrisankaran, and R. Town, (2003), "Bayesian Inference for Hospital Quality in a Selection Model," *Econometrica*, 71(4), 1215-1238.

Geweke, J., (1997), "Posterior Simulations in Econometrics," in *Advances in Economics and Econometrics: Theory and Applications*, Vol III, Kreps and Wallis (eds.), Cambridge University Press.

Gilks, W. S. Richardson and D. Spiegelhalter, (1996), *Markvo Chain Monte Carlo in Practice,* Chapman and Hall, New York, NY.

Goettler, J., and R. Shachar (2001), "Spatial Competition in the Network Television Industry," *RAND Journal of Economics*, Vol. 32(4), 624-656.

Lancaster, T., (2004), *An Introduction to Modern Bayesian Econometrics,* Blackwell Publishing, Malden, MA.

Rossi, P., R. McCulloch, and G. Allenby, (1996), "The Value of Purchasing History Data in Target Marketing," *Marketing Science*, Vol 15(4), 321-340.

Rossi, P., G. Allbeny, and R. McCulloch, (2005), *Bayesian Statistics and Marketing*, Wiley, Hoboken, NJ.

Sims, C., and H. Uhlig, (1991), "Understanding Unit Rotters: A Helicopter View," *Econometrica*, 59(6), 1591-1599.