"A Course in Applied Econometrics"

Lecture 13

Bayesian Inference

Guido Imbens

IRP Lectures, UW Madison, August 2008

Outline

1. Introduction
2. Basics
3. Bernstein-Von Mises Theorem
4. Markov-Chain-Monte-Carlo Methods
5. Example: Demand Models with Unobs Heterog in Prefer.
6. Example: Panel Data with Multiple Individual Specific Param.

7. Instrumental Variables with Many Instruments
8. Example: Binary Response with Endogenous Regressors
9. Example: Discrete Choice Models with Unobserved Choice Characteristics

1. Introduction

Formal Bayesian methods surprisingly rarely used in empirical work in economics.

Surprising, because they are attractive options in many settings, especially with many parameters (like random coefficient models), when large sample normal approximations are not accurate. (see examples below)

In cases where large sample normality does not hold, frequentist methods are sometimes awkward (e.g, confidence intervals that can be empty, such as in unit root or weak instrument cases).

Bayesian approach allows for conceptually straightforward way of dealing with unit-level heterogeneity in preferences/parameters.

Why are Bayesian methods not used more widely?

1. choice of methods does not matter (bernstein-von mises theorem)

2. difficulty in specifying prior distribution (not "objective")

3. need for fully parametric model

4. computational difficulties

2.A Basics: The General Case

Model:

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

As a function of the parameter this is called the likelihood function, and denoted by $\mathcal{L}(\theta|x)$.

A prior distribution for the parameters, $p(\theta)$.

The posterior distribution,

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

Note that, as a function of θ, the posterior is proportional to

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

2.B Example: The Normal Case

Suppose the conditional distribution of X given the parameter μ is $\mathcal{N}(\mu, 1)$.

Suppose the prior distribution for μ to be $\mathcal{N}(0, 100)$.

The posterior distribution is proportional to

$$f_{\mu|X}(\mu|x) \propto \exp\left(-\frac{1}{2}(x-\mu)^2\right)\cdot\exp\left(-\frac{1}{2\cdot 100}\mu^2\right)$$

$$= \exp -\frac{1}{2}(x^2 - 2x\mu + \mu^2 + \mu^2/100)$$

$$\propto \exp(-\frac{1}{2(100/101)}(\mu - (100/101)x)^2)$$

$$\sim \mathcal{N}(x\cdot 100/101, 100/101)$$

2.B The Normal Case with General Normal Prior Distribution

Model: $\mathcal{N}(\mu, \sigma^2)$
Prior distribution for μ is $\mathcal{N}(\mu_0, \tau^2)$.

Then the posterior distribution is:

$$f_{\mu|X}(\mu|x) \sim \mathcal{N}\left(\frac{x/\sigma^2 + \mu_0/\tau^2}{1/\sigma^2 + 1/\tau^2}, \frac{1}{1/\tau^2 + 1/\sigma^2)}\right).$$

The result is quite intuitive: the posterior mean is a weighted average of the prior mean μ_0 and the observation x with weights proportional to the precision, $1/\sigma^2$ for x and $1/\tau^2$ for μ_0:

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

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? We can 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 (flat prior). That would seem to capture pretty well the idea that a priori we are ignorant about μ.

This is not always easy to do. For example, a flat prior distribution is not always uninformative about particular functions of parameters.

2.C The Normal Case with Multiple Observations

N independent draws from $\mathcal{N}(\mu, \sigma^2)$, σ^2 known.

Prior distribution on μ is $\mathcal{N}(\mu_0, \tau^2)$.

The likelihood function is

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

Then

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

3.A Bernstein-Von Mises Theorem: normal example

When N is large

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

In large samples the prior does not matter.

Moreover, in a frequentist analysis, in large samples,

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

Bayesian probability and frequentiest confidence intervals agree:

$$\Pr\left(\mu \in \left[\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}, \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}\right] \middle| X_1, \ldots, X_N\right)$$

$$\approx \Pr\left(\mu \in \left[\overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}, \overline{X} - 1.96 \cdot \frac{\sigma}{\sqrt{N}}\right] \middle| \mu\right) \approx 0.95;$$

3.B Bernstein-Von Mises Theorem: general case

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

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 γ 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 $\widehat{\theta}_{mle}$ also has a limiting distribution with mean zero and variance σ^2:

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

The implication is that we can interpret confidence intervals as approximate probability intervals from a Bayesian perspective.

Specifically, let the 95% confidence interval be $[\widehat{\theta}_{ml} - 1.96 \cdot \widehat{\sigma}/\sqrt{N}, \widehat{\theta}_{ml} + 1.96 \cdot \widehat{\sigma}/\sqrt{N}]$. Then, approximately,

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

$$\longrightarrow 0.95.$$

3.C When Bernstein-Von Mises Fails

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 such settings one has to take a more principled stand whether one wants to make subjective probability statements, or frequentist claims.

4. Numerical Methods: Markov-Chain-Monte-Carlo

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}

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

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 distributionas $k \longrightarrow \infty$:

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

Then just pick a θ_0 and approximate the mean and standard deviation of the posterior distribution as

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

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

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

4.A 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

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

it may be easier to sample $\theta_{1,k+1}$ from the conditional distribution

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

and then sample $\theta_{2,k+1}$ from

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

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.B Data Augmentation

Suppose we are interested in estimating the parameters of a censored regression or Tobit model. There is a latent variable

$$Y_i^* = X_i'\beta + \varepsilon_i, \quad \varepsilon_i|X_i \sim \mathcal{N}(0,1)$$

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. 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

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

if observation i is truncated, where $\mathcal{TN}(\mu, \sigma^2, c)$ denotes a truncated normal distribution with mean μ, variance σ^2, and truncation point c (truncated from above).

Second, we draw a new value for the parameter, β_{k+1} given the data and given the (partly drawn) $\mathbf{Y}^*$:

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

4.C Metropolis Hastings

We are again interested in $p(\theta|\text{data})$. In this case $\mathcal{L}(\theta|\text{data})$ is assumed to be easy to evaluate. Draw a new candidate value for the chain from a candidate distribution $q(\theta|\theta_k, \text{data})$. We will either accept the new value with probability The probability that the new draw θ is accepted is

$$\rho(\theta_k, \theta) = \min\left(1, \frac{p(\theta|\text{data}) \cdot q(\theta_k|\theta, \text{data})}{p(\theta_k|\text{data}) \cdot q(\theta|\theta_k, \text{data})}\right),$$

so that

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

The optimal (typically infeasible) choice for the candidate distribution is

$$q^*(\theta|\theta_k, \text{data}) = p(\theta|\text{data}) \implies \rho(\theta_k, \theta) = 1$$

5. Example: Demand Models with Unobs Heterog in Prefer.

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.

Alternatively, they may choose to offer differential coupons based on consumer's fixed characteristics.

Taking this ever further, they could tailoring the coupon value to the evidence for price sensitivity contained in purchase patterns.

Need 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 hiearchical, 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 $Z_i'\Gamma$:

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

Now Gibbs sampling can be used to obtain draws from the posterior distribution of the β_i.

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.

6. Example: Panel Data with Multiple Individual Specific Param.

Chamberlain and Hirano are interested in deriving predictive distributions for earnings using longitudinal data, using the model

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

The second component in the model, V_{it}, is a first order autoregressive component,

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

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

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

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.

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

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

In their empirical application using data from the Panel Study of Income Dynamics (PSID), CH find strong evidence of heterogeneity in conditional variances.

quantiles of the predictive dist. of $1/\sqrt{h_i}$

	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
HS Dropouts (N=37)	0.06	0.08	0.11	0.16	0.27	0.49	0.79
HS Grads (N=100)	0.04	0.05	0.06	0.11	0.21	0.49	0.93
C Grads (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.

individual	sample std	0.90-0.10 quantile 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.

7. Example: Instrumental Variables with Many Instruments

Chamberlain and Imbens analyze the many instrument problem from a Bayesian perspective. 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.

A flat prior distribution may 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.

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.

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 rather than the $\sigma_\pi^2 = \infty$ implicit in TSLS. 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 implying 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.

8. Example: Binary Response with Endogenous Regressors

Geweke, Gowrisankaran, and Town 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 characteristics Z_i, latent mortality is

$$Y_i^* = \sum_{j=1}^{114} C_{ij}\beta_j + Z_i'\gamma + \epsilon_i,$$

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

$$\epsilon_i = \sum_{j=1}^{114} \eta_{ij} \cdot \delta_j + \zeta_i,$$

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.

9. Example: 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, 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 mcmc 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.