Econ 226: Bayesian and Numerical Methods

- Course requirements: one exam (based primarily on lectures) and paper proposal
- Slides: sometimes, not always, check web page night before
- Office hours: Mondays 9:30-10:30 a.m.
- Theme: Bayesian econometrics and applications of numerical Bayesian methods

Why Bayesian?

(1) Allows us to incorporate information in addition to that in the sample
   (a) VARs
   (b) short time series, measurement error
   (c) lag lengths, nonstationarity
(2) Empirical use of DSGEs
(3) Clean solution to many otherwise tricky questions
(4) Analyze models that are intractable using frequentist methods

I. Bayesian Econometrics

A. Introduction

Consider simple model:

$y_t = \mu + \varepsilon_t \qquad t = 1, 2, \ldots, T$

$\varepsilon_t \sim N(0, \sigma^2)$ i.i.d.

σ^2 known, want to estimate μ

Estimate (MLE = OLS = GMM):

$\hat{\mu} = T^{-1} \sum_{t=1}^{T} y_t$

$\hat{\mu} \sim N(\mu, \sigma^2/T)$

Beginning student wants to say, "there is a 95% probability that μ is in the interval $\hat{\mu} \pm 1.96\sigma/\sqrt{T}$"

Frequentist statistician says:

"No, no, no! μ is the true value. It either equals 5 or it doesn't. There is no probability statement about μ.

"What is true is that if we use this procedure to construct an interval in thousands of different samples, in 95% of those samples, our interval will contain the true μ."

Suppose we observe a sample mean of 5 and know that $\sigma/\sqrt{T} = 1$, and then ask the frequentist statistician:

"Do you know the true μ?"

"No."

"Choose between these options. Option A: I give you $5 now. Option B: I give you $10 if the true μ is in the interval between 2.0 and 3.5."

"I'll take the $5, thank you."

"How about these? Option A: I give you $5 now. Option B: I give you $10 if the true μ is between -5.0 and +10.0."

"OK, I'll take option B."

"Option A: I generate a uniform number between 0 and 1. If the number is less than π, I give you $5. Option B: I give you $5 if the true μ is in the interval (2.0,4.5). The value of π is 0.2"

"Option B."

"How about if $\pi = 0.8$?"

"Option A."

If the statistician's choices among such comparisons satisfy certain axioms of rationality, then there will exist a unique π^* such that he chooses Option A whenever $\pi > \pi^*$ and Option B whenever $\pi < \pi^*$. We might interpret this π^* as the statistician's (subjective) probability that μ is in the interval $(2.0, 4.0)$.

Bayesian idea: before seeing the data $(y_1, y_2, \ldots, y_T)$, the statistician had some subjective probability beliefs about the value of μ, called the "prior distribution."

Suppose we represent these beliefs with a probability distribution, $p(\mu)$, called the "prior distribution."
For example, $\mu \sim N(m, \tau^2)$.

$$p(\mu) = \frac{1}{\sqrt{2\pi}\,\tau} \exp[-(\mu - m)^2/(2\tau^2)]$$

m represents our "best guess" as to the value of μ before seeing data

τ^2 represents our confidence in this guess– small τ, very confident

We think of the usual likelihood function as the probability of the data given fixed values for μ and σ:

$$p(\mathbf{y}|\mu;\sigma) = \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{\frac{-\sum_{t=1}^{T}(y_t-\mu)^2}{2\sigma^2}\right\}$$

We can further think of the joint probability distribution of $\mathbf{y}$ and μ, characterizing our joint uncertainty about parameters and data:

$$p(\mathbf{y},\mu;\sigma) = p(\mathbf{y}|\mu;\sigma) \cdot p(\mu)$$

$$= \frac{1}{(2\pi)^{(T+1)/2}\sigma^T\tau} \exp\left\{\frac{-(\mu-m)^2}{2\tau^2} - \frac{\sum_{t=1}^{T}(y_t-\mu)^2}{2\sigma^2}\right\}$$

The goal of statistical analysis is to characterize our subjective beliefs about μ after having seen the data, called the "posterior distribution":

$$p(\mu|\mathbf{y};\sigma) = \frac{p(\mathbf{y}|\mu;\sigma)p(\mu)}{p(\mathbf{y})} = \frac{p(\mathbf{y}|\mu;\sigma)p(\mu)}{\int p(\mathbf{y}|\mu;\sigma)p(\mu)\,d\mu}$$

One way to find this posterior distribution is by brute force (integrating and dividing).

An easier way to come up with the identical answer is to factor the joint density into a component that depends on μ and a component that does not depend on μ:

$$p(\mathbf{y},\mu;\sigma) = \frac{1}{(2\pi)^{(T+1)/2}\sigma^T\tau}\exp\left\{\frac{-(\mu-m)^2}{2\tau^2} - \frac{\sum_{t=1}^T (y_t-\mu)^2}{2\sigma^2}\right\}$$

$$\propto \exp\left\{-\frac{\mu^2-2\mu m}{2\tau^2} - \frac{T\mu^2-2\mu\sum_{t=1}^T y_t}{2\sigma^2}\right\}$$

$$p(\mu|\mathbf{y};\sigma) = \frac{p(\mathbf{y}|\mu;\sigma)p(\mu)}{\int p(\mathbf{y}|\mu;\sigma)p(\mu)\,d\mu}$$

The expression in the denominator, $\int p(\mathbf{y}|\mu;\sigma)p(\mu)\,d\mu$, does not depend on μ, and is really just contributing the constant (with respect to μ) that we have to divide $p(\mathbf{y},\mu;\sigma)$ by to get something that is a proper density (with respect to μ)

If we knew what that something has to be (from recognizing the kernel as part of a known density), we can jump immediately to the result of integrating and dividing.

$$p(\mathbf{y},\mu;\sigma) \propto \exp\left\{-\frac{\mu^2-2\mu m}{2\tau^2} - \frac{T\mu^2-2\mu\sum_{t=1}^{T} y_t}{2\sigma^2}\right\}$$

$$\propto \exp\left\{-\frac{(\sigma^2/T)(\mu^2-2\mu m)}{2(\sigma^2/T)\tau^2} - \frac{\tau^2(\mu^2-2\mu\overline{y})}{2(\sigma^2/T)\tau^2}\right\}$$

$$\propto \exp\left\{-\frac{[(\sigma^2/T)+\tau^2]\left[\mu^2-\frac{2\mu m(\sigma^2/T)}{(\sigma^2/T)+\tau^2}-\frac{2\mu\overline{y}\tau^2}{(\sigma^2/T)+\tau^2}\right]}{2(\sigma^2/T)\tau^2}\right\}$$

$$\propto \exp\left\{-\frac{[(\sigma^2/T)+\tau^2][\mu^2-2\mu m^*]}{2(\sigma^2/T)\tau^2}\right\}$$

for $m^* = \left(\frac{m(\sigma^2/T)}{(\sigma^2/T)+\tau^2} + \frac{\overline{y}\tau^2}{(\sigma^2/T)+\tau^2}\right), \tau^{*2} = \frac{\tau^2\sigma^2/T}{(\sigma^2/T)+\tau^2}$

$$p(\mathbf{y}, \mu; \sigma) \propto \exp\left\{-\frac{(\mu^2 - 2\mu m^* + m^{*2})}{2\tau^{*2}}\right\}$$

$$\propto \frac{1}{\sqrt{2\pi}\,\tau^*} \exp\left\{-\frac{(\mu - m^*)^2}{2\tau^{*2}}\right\}$$

which is the $N(m^*, \tau^{*2})$ density, which integrates to unity (with respect to μ).

Therefore

$$p(\mu|\mathbf{y}; \sigma) = \frac{p(\mathbf{y}, \mu; \sigma)}{\int p(\mathbf{y}, \mu; \sigma)\, d\mu}$$

$$= \frac{1}{\sqrt{2\pi}\,\tau^*} \exp\left\{-\frac{(\mu - m^*)^2}{2\tau^{*2}}\right\}$$

i.e., $\mu|\mathbf{y} \sim N(m^*, \tau^{*2})$

$$m^* = \frac{m(\sigma^2/T)}{(\sigma^2/T)+\tau^2} + \frac{\overline{y}\,\tau^2}{(\sigma^2/T)+\tau^2}$$

weighted average of m (what we thought without seeing data) and $\overline{y}$ (value suggested by the data)

$\tau \to \infty \Rightarrow$ prior information worthless (called “diffuse” prior)

$\Rightarrow m^* \to \overline{y}$

$m^* = \frac{m(\sigma^2/T)}{(\sigma^2/T)+\tau^2} + \frac{\bar{y}\tau^2}{(\sigma^2/T)+\tau^2}$

$\tau \to 0 \Rightarrow$ absolutely certain before seeing data

$\Rightarrow m^* \to m$

nothing in data could change my mind

$m^* = \frac{m(\sigma^2/T)}{(\sigma^2/T)+\tau^2} + \frac{\bar{y}\tau^2}{(\sigma^2/T)+\tau^2}$

Given σ^2 and τ^2, as $T \to \infty$,

$m^* \longrightarrow \bar{y}$

eventually data overwhelm any reasonable prior

$\tau^{*2} = \frac{\tau^2\sigma^2/T}{(\sigma^2/T)+\tau^2}$

summarizes confidence in posterior conclusion

$T \to \infty \Rightarrow \tau^{*2} \to 0$

as accumulate data, become more confident

Diffuse prior:

$\mu|\mathbf{y} \sim N(m^*, \tau^{*2})$

$\sim N(\overline{y}, \sigma^2/T)$

Bayesian statistician: “Having seen the data, there is a 95% probability that μ is in the interval $\overline{y} \pm 1.96\sigma/\sqrt{T}$”

I. Bayesian econometrics

A. Introduction
B. Bayesian inference in the univariate regression model

Consider

$$y_t = \mathbf{x}_t'\boldsymbol{\beta} + \varepsilon_t$$

$$\varepsilon_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1, y_{t-1}, y_{t-2}, \ldots, y_1 \sim N(0, \sigma^2)$$

likelihood:

$$p(\mathbf{y}|\boldsymbol{\beta}, \sigma^2) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t, \mathbf{x}_{t-1}, \ldots, \mathbf{x}_1)$$

$$= \frac{1}{(2\pi\sigma^2)^{T/2}} \exp\left\{ -\frac{\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2} \right\}$$

If $Z_i \sim N(0,\tau^2)$ and $W = \sum_{i=1}^{N} Z_i^2$

then $W \sim \Gamma(N,\lambda)$ for $\lambda = \tau^{-2}$:

$p(w) = [\Gamma(N/2)]^{-1}(\lambda/2)^{N/2} w^{[(N/2)-1]}$

$\exp[-\lambda w/2]$

$\Gamma(\alpha) = \int_0^\infty x^{\alpha-1}e^{-x}dx$

(e.g., $\Gamma(\alpha) = (\alpha-1)!$ for $\alpha \in \{1,2,\ldots\}$

Prior distribution: The inverse of the variance (σ^{-2}, also called the "precision") is distributed $\Gamma(N,\lambda)$

$p(\sigma^{-2}) = [\Gamma(N/2)]^{-1}(\lambda/2)^{N/2}(\sigma^{-2})^{[(N/2)-1]}$

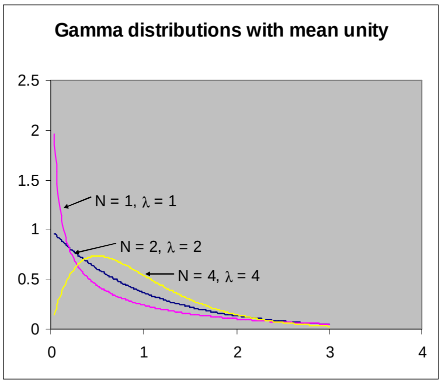
$\exp[-\lambda\sigma^{-2}/2]$

e.g., our prior is equivalent to earlier having observed N observations with sum of squared residuals λ

Why use this prior?

1) $p(\sigma^2) = 0$ for $\sigma^2 < 0$

2) flexible family (different shapes)



3) It is the "natural conjugate prior" given the specified $N(\mathbf{x}_t'\boldsymbol{\beta}, \sigma^2)$ likelihood, meaning if prior is $\sigma^{-2} \sim \Gamma(N, \lambda)$, then posterior turns out to be $\sigma^{-2}|\mathbf{y} \sim \Gamma(N^*, \lambda^*)$

a) if our prior was based on earlier data analysis, would have this form

b) makes analytical treatment of problem tractable

Prior distribution for the $(k \times 1)$ vector $\boldsymbol{\beta}$ conditional on σ^{-2}:

$$\boldsymbol{\beta}|\sigma^{-2} \sim N(\mathbf{m}, \sigma^2\mathbf{M})$$

$$p(\boldsymbol{\beta}|\sigma^2) = (2\pi\sigma^2)^{-k/2}|\mathbf{M}|^{-1/2} \exp\left\{\left[-\tfrac{1}{2\sigma^2}\right](\boldsymbol{\beta}-\mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta}-\mathbf{m})\right\}$$

prior guess for $\boldsymbol{\beta} = \mathbf{m}$

much uncertainty about this guess: diagonal elements of $\mathbf{M}$ large

$$p(\boldsymbol{\beta}, \sigma^{-2}|\mathbf{y}) \propto p(\mathbf{y}, \boldsymbol{\beta}, \sigma^{-2})$$

$$\propto (\sigma^{-2})^{T/2} \exp\left\{-\frac{\sum_{t=1}^{T}(y_t - \mathbf{x}_t'\boldsymbol{\beta})^2}{2\sigma^2}\right\}$$

$$(\sigma^{-2})^{[(N/2)-1]} \exp[-\lambda\sigma^{-2}/2]$$

$$(\sigma^{-2})^{k/2} \exp\left\{\left[-\tfrac{1}{2\sigma^2}\right](\boldsymbol{\beta} - \mathbf{m})'\mathbf{M}^{-1}(\boldsymbol{\beta} - \mathbf{m})\right\}$$

$$\sigma^{-2}|\mathbf{y} \sim \Gamma(N^*, \lambda^*)$$

$$N^* = N + T$$

$$\lambda^* = \lambda + \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\mathbf{b})^2$$

$$+ (\mathbf{b} - \mathbf{m})'\tilde{\mathbf{M}}(\mathbf{b} - \mathbf{m})$$

$$\mathbf{b} = \left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t y_t\right)$$

$$\tilde{\mathbf{M}} = \mathbf{M}^{-1}\left(\mathbf{M}^{-1} + \sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)$$

$$\boldsymbol{\beta}|\mathbf{y}, \sigma^{-2} \sim N(\mathbf{m}^*, \sigma^2\mathbf{M}^*)$$

$$\mathbf{m}^* = \mathbf{M}^*\left(\mathbf{M}^{-1}\mathbf{m} + \sum_{t=1}^{T}\mathbf{x}_t y_t\right)$$

$$\mathbf{M}^* = \left(\mathbf{M}^{-1} + \sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}$$

diffuse prior: $\mathbf{M} \to \infty \cdot \mathbf{I}_k$

$$\Rightarrow \mathbf{M}^* \to \left(\sum \mathbf{x}_t\mathbf{x}_t'\right)^{-1}$$

$$\Rightarrow \mathbf{m}^* \to \left(\sum \mathbf{x}_t\mathbf{x}_t'\right)^{-1}\left(\sum \mathbf{x}_t y_t\right)$$

= usual OLS formulas

Can also show

$\boldsymbol{\beta}|\mathbf{y} \sim$ Student t with $N+T$ degrees of freedom, mean $\mathbf{m}^*$, and scale matrix $(\lambda^*/N^*)\mathbf{M}^*$ for

$$\lambda^* = \lambda + \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2 + (\hat{\boldsymbol{\beta}} - \mathbf{m})'\widetilde{\mathbf{M}}(\hat{\boldsymbol{\beta}} - \mathbf{m})$$

$$\widetilde{\mathbf{M}} = \mathbf{M}^{-1}\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t' + \mathbf{M}^{-1}\right)^{-1}\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'$$

Diffuse prior:

$\lambda = N = 0, \mathbf{M}^{-1} = \mathbf{0}$

$\boldsymbol{\beta}|\mathbf{y} \sim$ Student t with T degrees of freedom, mean $\hat{\boldsymbol{\beta}}$, and scale matrix $(\hat{\lambda}/T)\left(\sum_{t=1}^{T}\mathbf{x}_t\mathbf{x}_t'\right)^{-1}$ for

$$\hat{\lambda} = \sum_{t=1}^{T}(y_t - \mathbf{x}_t'\hat{\boldsymbol{\beta}})^2$$

Same as usual OLS results, except with T instead of $T-k$ for degrees of freedom and denominator of $\hat{s}^2$

Classical perspective:

fixed regressors: Student t (with $T-k$ d.f.) is exact small-sample distribution

univariate AR(p): Student t is only asymptotic distribution

Bayesian perspective:

Student t is exact small sample result regardless

Classical perspective assumes a single true β_0, integrates over realizations of $\mathbf{y}$:

$$E(\hat{\beta}) = \int_{\Re^T} \hat{\beta}(\mathbf{y}) f_{\beta_0}(\mathbf{y}) d\mathbf{y}$$

Bayesian perspective conditions on the data, integrates over all values of β:

$$E(\beta|\mathbf{y}) = \int_{\Re^k} \beta p(\beta|\mathbf{y}) d\beta$$

$$\beta|\mathbf{y}, \sigma^{-2} \sim N(\mathbf{m}^*, \sigma^2 \mathbf{M}^*)$$

$$\mathbf{m}^* = \mathbf{M}^*\left(\mathbf{M}^{-1}\mathbf{m} + \sum_{t=1}^{T} \mathbf{x}_t y_t\right)$$

$$\mathbf{M}^* = \left(\mathbf{M}^{-1} + \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1}$$

dogmatic prior: $\mathbf{M} \to 0 \cdot \mathbf{I}_k$

$\Rightarrow \mathbf{m}^* \to \mathbf{m}$

$\Rightarrow \mathbf{M}^* \to \mathbf{0}$

posterior = prior

$$\beta|\mathbf{y}, \sigma^{-2} \sim N(\mathbf{m}^*, \sigma^2 \mathbf{M}^*)$$

$$\mathbf{m}^* = \mathbf{M}^*\left(\mathbf{M}^{-1}\mathbf{m} + \sum_{t=1}^{T} \mathbf{x}_t y_t\right)$$

$$\mathbf{M}^* = \left(\mathbf{M}^{-1} + \sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t'\right)^{-1}$$

in general: $\mathbf{m}^*$ is weighted average of $\mathbf{m}$ and $\hat{\beta}$, where weights depend on confidence in prior ($\mathbf{M}$) and strength of evidence from data ($\sum \mathbf{x}_t \mathbf{x}_t'$)

Another way to interpret prior:
Suppose I had observed an earlier sample of $\tilde{T}$ observations:

$$\{\tilde{y}_{\tilde{t}}, \tilde{\mathbf{x}}_{\tilde{t}}\}_{\tilde{t}=1}^{\tilde{T}}$$

which were independent of the current observed sample:

$$\{y_t, \mathbf{x}_t\}_{t=1}^{T}$$

Then my OLS estimate based on all information would be

$$\boldsymbol{\beta}^* = \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t y_t + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{y}_{\tilde{t}}\right)$$

with variance (given σ^2) of

$$\mathrm{Var}(\boldsymbol{\beta}^*) = \sigma^2 \left(\sum_{t=1}^{T} \mathbf{x}_t \mathbf{x}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1}$$

Let $\mathbf{m}$ be the OLS estimate based on the prior sample alone,

$$\mathbf{m} = \left(\sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1} \left(\sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{y}_{\tilde{t}}\right)$$

and let $\sigma^2 \mathbf{M}$ denote its variance:

$$\mathbf{M} = \left(\sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}} \tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1}$$

$$\boldsymbol{\beta}^* = \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}}\tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t y_t + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}}\tilde{y}_{\tilde{t}}\right) = \left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' + \mathbf{M}^{-1}\right)^{-1} \left(\sum_{t=1}^{T} \mathbf{x}_t y_t + \mathbf{M}^{-1}\mathbf{m}\right)$$

identical to formula for posterior mean $\mathbf{m}^*$

$$\mathrm{Var}(\boldsymbol{\beta}^*) = \sigma^2\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' + \sum_{\tilde{t}=1}^{\tilde{T}} \tilde{\mathbf{x}}_{\tilde{t}}\tilde{\mathbf{x}}_{\tilde{t}}'\right)^{-1} = \sigma^2\left(\sum_{t=1}^{T} \mathbf{x}_t\mathbf{x}_t' + \mathbf{M}^{-1}\right)^{-1} = \sigma^2\mathbf{M}^*$$

for $\mathbf{M}^*$ the posterior variance given earlier

So what priors would we believe?

- Fama: stock prices are random walk
- Hall: consumption is random walk
- Mankiw: marginal tax rates are random walk

Equation of a VAR:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p}$$
$$+\boldsymbol{\varphi}_1' \mathbf{x}_{t-1} + \boldsymbol{\varphi}_2' \mathbf{x}_{t-2} + \cdots + \boldsymbol{\varphi}_p' \mathbf{x}_{t-p} + \varepsilon_t$$

We expect $\phi_1 = 1$ and all other coefficients are zero

$$\mathbf{m} = (0, 1, 0, \ldots, 0)$$

Have more confidence in these values (diagonal elements of $\mathbf{M}$ smaller) for other variables ($\boldsymbol{\varphi}_j$ versus ϕ_j) and for higher-order lags (j bigger)

= “Minnesota prior”

Is prior information good?

a) Is random walk good approximation?

b) Shrinkage often improves forecasts

c) Unavoidable trade-off: objectivity versus accuracy