14.384 Time Series Analysis, Fall 2007
Professor Anna Mikusheva
Paul Schrimpf, scribe
Novemeber 29, 2007
revised December 3, 2009

Lecture 23-24

Intro to Bayes approach. Reasons to be Bayesian

Many ideas are borrowed from Lecture notes written by Frank Schorfheide.
Bayesian models has two pieces:

- A parametric model, giving a distribution, $f(\mathcal{Y}_T|\theta)$, for the data given parameters
- A prior distribution for the parameters, $p(\theta)$

From these, we can form the joint distribution of the data and parameters,

$$p(\mathcal{Y}_T, \theta) = f(\mathcal{Y}_T|\theta)p(\theta)$$

and the marginal distribution of the data

$$p(\mathcal{Y}_T) = \int f(\mathcal{Y}_T|\theta)p(\theta)d\theta$$

Finally, using Baye's rule, we can form the posterior distribution of the parameters given the data

$$p(\theta|\mathcal{Y}_T) = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\mathcal{Y}_T)}$$

One can make inferences based on the posterior distribution. For example, we can report the mode (or the mean) as a parameter estimate. Any set of posterior measure biggen than $1-\alpha$ is called an $1-\alpha$ credible set. Hypotheses can be tested based on posterior odds.

Differences between Bayesian and Frequentist Approaches

Frequentist

- θ is fixed, but unknown
- Uncertainty comes from sampling uncertainty. That is, from the fact that we can get different samples.
- All probabilistic statements are statements about sampling uncertainty. For example,
  - $E_\theta \hat{\theta}(\mathcal{Y}_T) = \theta$ (unbiasedness) means in average over all possible repeated samples, one receives the true value
  - $P_\theta\{\theta \in C(\mathcal{Y}_T)\} = 1-\alpha$ coverage probability of confidence sets is a statement about the ratio (in repeated samples) of sets containing θ. Once we observe a sample, θ is either in the set or not; there is no probability, after realization of a sample. The coverage probability is a statement about ex-ante probability.

Bayesian

- θ is random
- $\mathcal{Y}_T$ is treated as fixed after observed
- uncertainty = "beliefs" about θ
- All probabilistic statements are about uncertainty about θ.
  - $P_\theta\{\theta \in C(\mathcal{Y}_T)\} = 1 - \alpha$ coverage probability is the probability that θ is in the set.

Reasons to be Bayesian

Reason 1 – Philosophical

Example 1. Two observations, $y_1, y_2 \sim$ iid.

$$P_\theta(y_i = \theta - 1) = \frac{1}{2} = P_\theta(y_i = \theta + 1)$$

Consider a confidence set:

$$C(y_1, y_2) = \begin{cases} \frac{y_1+y_2}{2} & y_1 \neq y_2 \\ y_1 - 1 & y_1 = y_2 \end{cases}$$

If we observe $y_1 \neq y_2$, which will happen 1/2 the time, we know $\theta = C(y_1, y_2)$. Otherwise, we have a probability of 1/2 that $\theta = C(y_1, y_2)$. From a frequentist perspective, then the coverage of this set is $1/2 + 1/2 * 1/2 = 75\%$. Now, suppose we observe $y_1 \neq y_2$. Then we know θ with certainty. Why would we then report a coverage of 75% (ex-ante coverage) rather than the ex-post accuracy of 100%? Frequentists average probabilities over all situations that may have been realized, but were not. Bayesians are conditioning on the realization. As this example shows, conditioning on observation may be justified.

This example might appear artificial. However, especially in time series, there are fundamental reasons to be Bayesian. Perhaps the frequentist perspective makes sense in a cross section. In a cross section, we can imagine taking different samples and repeating our "experiment" (idea of repeated samples). However, in time series we often have only one realization, and it is difficult to imagine where we would obtain another sample. For example, if our data is U.S. inflation, then we would need another world to get another sample.

Bayes is based on *Conditional principle*: if experiment is selected by some random mechanism independent of θ, then only the experiment actually performed is relevant, not the experiments that may have been performed.

Another philosophical reason to be Bayesian is that we do have prior beliefs about parameters and we should incorporate them in a coherent way. Some people criticize Bayesians for imposing too much parametric structure and prior beliefs. Bayesians argue that even frequentists implicitly impose priors by choosing which models to estimate and which results to report. At least the role of priors is more apparent in Bayesian econometrics.

Conjugate Priors

When a prior and a posterior are in the same family of distributions, then the prior is called the conjugate prior for the distribution of the data. There are a handful of well behaved cases, which are very convenient and easier to work with.

Example 2. *OLS* The model is

$$y_t = x_t\theta + u_t$$

with $u_t \sim iidN(0,1)$. In matrix form we'll write $Y = X\theta + U$. The distribution of the data is

$$f(Y|X,\theta) = (2\pi)^{-T/2}\exp\left(-\frac{1}{2}(Y-X\theta)'(Y-X\theta)\right)$$

If we choose a normal prior, $\theta \sim N(0,\tau^2 I_k)$,

$$p(\theta) = (2\pi\tau^2)^{-k/2}\exp\left(\frac{-1}{2\tau}\theta'\theta\right)$$

then the posterior is

$$\begin{aligned} p(\theta|Y,X) &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'X'X\theta + \frac{1}{\tau^2}\theta'\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[-Y'X\theta - \theta'X'Y + \theta'(X'X + \frac{I_k}{\tau^2})\theta\right]\right) \\ &\propto \exp\left(-\frac{1}{2}\left[\left(\theta - (X'X+\frac{I_k}{\tau^2})^{-1}X'Y\right)'(X'X+\frac{I_k}{\tau^2})\left(\theta - (X'X+\frac{I_k}{\tau^2})^{-1}X'Y\right)\right]\right) \end{aligned}$$

so $\theta|Y,X \sim N(\tilde{\theta},\tilde{\Sigma})$ with

$$\begin{aligned} \tilde{\theta} =& (X'X+\frac{I_k}{\tau^2})^{-1}X'Y \\ \tilde{\Sigma} =& (X'X+\frac{I_k}{\tau^2})^{-1} \end{aligned}$$

Also, we see that as $\tau \to \infty$ (uninformative prior), $\tilde{\theta} \to (X'X)^{-1}X'Y = \hat{\theta}^{ML}$, and as $\tau \to 0$, $\tilde{\theta} \to 0$, the prior dominates. Furthermore, if we fix τ and $T \to \infty$ with $\frac{X'X}{T} \to Q_{XX}$, then $\tilde{\theta} \to \theta_0$, the frequentist limit. So, prior vanishes asymptotically. This result is more general.

Reason 2- priors vanish asymptotically

The prior vanishes asymptotically. All inference in "clean" situations asymptotically converge to frequentists'. This claim is based on the following theorem:

Theorem 3. *(Geweke, p.93) Suppose*

1. *Prior is absolutely continuous wrt to Lebesgue measure and prior puts positive probability on all sets with positive Lebesgue measure*
2. *Uniform convergence of likelihood $\frac{1}{T}\log f(\mathcal{Y}_T|\theta) \to^{a.s.} l(\theta)$ uniformly in θ*
3. *$l(\theta)$ is continuous and has a unique maximum at θ^**

Then for any open neighborhood, $\varepsilon(\theta^*)$,

$$\lim_{T\to\infty} P(\theta \in \varepsilon(\theta^*)|\mathcal{Y}_T) = 1 \ a.s.$$

Theorem says that the posterior concentrates around the asymptotic limit of frequentist MLE, and in any "reasonable" situation Baeys estimate in large samples will be close to MLE. There's a similar theorem that shows asymptotic normality of the Bayesian estimator. These two theorems sort of say what happens when you use Bayesian methods in a frequentist world. One interpretation is that the prior vanishes asymptotically. However, you should be cautious about this theorem!
Cautions:

- The theorem is about asymptotics. However, the prior can influence inferences in finite samples.
- Condition 3 is an identification condition. If you are not identified, then where the Bayesian estimator converges depends on your prior.
- Prior should not restrict parameter space (condition 1)
- Condition 2 is like a LLN, it may not be satisfied with non-stationarity

Reason 3 – Decision Theory

All Bayes rules possess an optimality property in frequentists' world- they are admissible. And under some conditions all admissible rules are Bayes. For explaining what it means we have to introduce some new concepts.

Suppose we have a loss function $\mathcal{L}(a, \theta)$, where θ is a parameter and a is some action that we want to choose. For example, if we just want to estimate θ, we might have $a = \hat{\theta}$ and $\mathcal{L}(a, \theta) = (a - \theta)^2$. Our goal is to come up with a decision rule $a(\mathcal{Y}_T)$ that depends on the sample $\mathcal{Y}_T$, and give a small loss. Let our expected loss(called *risk*) for a given value of θ be

$$R_a(\theta) = E_\theta \mathcal{L}(a(\mathcal{Y}_T), \theta)$$

Note, that it is frequentist notion (expectation is taken over repeated samples)!!! In example above the risk is MSE. We would like $a(\mathcal{Y}_T)$ to minimize our expected loss. However, in general, the solution will depend on θ.

Definition 4. A decision rule, $a()$, is *admissible* if there exists no $\tilde{a}()$ such that $R_a(\theta) \geq R_{\tilde{a}}(\theta) \forall \theta$ with strict inequality for some θ_0.

Definition 5. A *Bayesian Decision Rule* minimizes a weighted risk (with weights $p(\theta)$):

$$\begin{aligned}
\min_a \int R_a(\theta) p(\theta) d\theta &= \min_a \int E_\theta \mathcal{L}(a(\mathcal{Y}_T), \theta) p(\theta) d\theta \\
&= \min_a \int \int \mathcal{L}(a(\mathcal{Y}_T), \theta) f(\mathcal{Y}_T | \theta) p(\theta) d\mathcal{Y}_T d\theta \\
&= \min_a \int \left[\int \mathcal{L}(a(\mathcal{Y}_T), \theta) p(\theta | \mathcal{Y}_T) d\theta \right] p(\mathcal{Y}_T) d\mathcal{Y}_T \\
&= \int \left[\min_a \int \mathcal{L}(a(\mathcal{Y}_T), \theta) p(\theta | \mathcal{Y}_T) d\theta \right] p(\mathcal{Y}_T) d\mathcal{Y}_T
\end{aligned}$$

So, Bayes rule for each realization $\mathcal{Y}_T$ solves for minimum of the posterior risk!

Theorem 6. *All Bayesian decision rules are admissible. Also, under some conditions, all admissible decision rules are Bayesian.*

Reason 4 – Nuisance Parameters

Let $\omega = h(\theta)$. Let $C(\mathcal{Y}_T)$ be a set such that $P(\omega \in C(\mathcal{Y}_T) | \mathcal{Y}_T) = 1 - \alpha$. It is very easy to go from $p(\theta | \mathcal{Y}_T)$ to $p(\omega | \mathcal{Y}_T)$. For example, suppose $\theta = (\theta_1, \theta_2)$ and $\omega = \theta_1$, then

$$p(\theta_1 | \mathcal{Y}_T) = \int p(\theta_1, \theta_2 | \mathcal{Y}_T) d\theta_2$$

This example is especially relevant because there are many examples in econometrics where we want to eliminate nuisance parameters. Here, θ_2 would be the nuisance parameters. The Bayesian approach makes it very easy to deal with the nuisance parameters. Whereas in the frequentist world, nuisance parameters are an extremely difficult problem.

Reason 5 – Easier to Implement

In some cases it can be easier to compute Bayesian estimates than frequentist ones. For example, in the last lecture, we saw how it can be difficult to estiate DSGE models by MLE. The main difficulty is in multidimensional optimization. In some cases, it can be easier to estimate these models using Bayesian methods and MCMC.

Reason 6 – Not Fully Identified, Priors add Identification

This is probably a bad reason to be Bayesian. If your model is not identified, then your estimates will be strongly influenced by your prior.

Bayes estimates, tests and sets.

Point Estimation

There are several Byes estimates depending on your loss function.

Our action is to choose a parameter from the parameter set $a \in \mathcal{A} = \Theta$. There are a number of loss functions we could use:

1. Quadratic loss function:
$$\mathcal{L}(\delta, \theta) = (\delta - \theta)' Q (\delta - \theta)$$
For some positive definite Q. If $Q = I$, then the risk is MSE. The corresponding Byes rule will be posterior mean $\hat{\delta} = E(\theta|\mathcal{Y}_T)$.

2. The loss function could be
$$\mathcal{L}(\delta, \theta) = \begin{cases} 1 & \delta \neq \theta \\ 0 & \delta = \theta \end{cases}$$
. If the parameter space is finite, then the corresponding risk is the probability of choosing a wrong value of θ. The optimal decision rule is then
$$\delta(\mathcal{Y}_T) = \arg\max_{\theta'} P(\theta = \theta'|\mathcal{Y}_T), \text{ which is the mode}$$
In the continuous case, $\forall \epsilon > 0$,
$$\mathcal{L}(\delta, \theta; \epsilon) = 1 - \mathbf{1}_{\{\theta \in \mathcal{N}_\epsilon(\delta)\}}$$
Let δ_ϵ denote the optimal decision rule for this loss function. Then $\lim_{\epsilon \to 0} \delta_\epsilon$ = the mode of the posterior distribution.

3. Check function:
$$\mathcal{L}(\delta, \theta) = (1 - q)(\delta - \theta)\mathbf{1}_{\{\theta < \delta\}} + q(\theta - \delta)\mathbf{1}_{\theta > \delta}$$
Then $\hat{\delta}$ is the q-th quantile of the posterior distribution of θ. For $q = 1/2$ we have the median.

Testing

The null hypothesis is $H_0 : \theta \in \Theta_0$, and the alternative is $H_1 : \theta \in \Theta_1$. Our action space is $\mathcal{A} = \{0 = \text{reject}, 1 = \text{accept}\}$. In a frequentists setting, the null and alternative hypotheses are not treated equally. The null is accepted unless there is strong evidence against it. In a Bayesian setting, the way the null and alternative are treated depends on the loss function. Consider the following loss function:

$$\mathcal{L}(\delta, \theta) = \begin{cases} 0 & \text{if correct} \\ a_1 & \delta = 0, \theta \in \Theta_0 \text{ (type 1 error)} \\ a_2 & \delta = 1, \theta \in \Theta_1 \text{ (type 2 error)} \end{cases}$$

Then the optimal decision rule is

$$\delta(\mathcal{Y}_T) = \begin{cases} 1 & \text{if } P(\theta \in \Theta_0|\mathcal{Y}_T) \geq \frac{a_2}{a_1+a_2} \\ 0 & \text{otherwise} \end{cases}$$

That is, we accept if the expected posterior loss from a type 1 error is bigger than the expected loss of a type 2 error

$$\begin{aligned} a_1 P(\theta \in \Theta_0|\mathcal{Y}_T) \geq & a_2 P(\theta \in \Theta_1|\mathcal{Y}_T) \\ \frac{P(\theta \in \Theta_0|\mathcal{Y}_T)}{P(\theta \in \Theta_1|\mathcal{Y}_T)} \geq & \frac{a_2}{a_1} \end{aligned}$$

from which we see that the null and alternative are treated symmetrically (one could easily relabel the null and alternative).

Example 7. $\Theta = \{0, 1\}$, we observe $y \in \{0, 1, 2\}$. The distribution of y given θ is:

y	0	1	2
$P_{\theta=0}$	0.89	0.04	0.07
$P_{\theta=1}$	0.959	0.04	0.001

In a frequentist world, if we observe $y = 1$ and test $H_0 : \theta = 1$, we will reject because the $P(y \geq 1|\theta = 1) < 0.05$. In a Bayesian, world we condition our test on the observed value of y and $y = 1$ is not informative about whether $\theta = 1$ or not (the posterior equals the prior, $p(\theta = 1|y = 1) = p(\theta = 1)$).

OLS

As in the previous lecture, the model is:

$$y_t = \theta X + U$$

with $U \sim iidN$. The prior is $\theta \sim N(0, \tau^2 I_k)$. Then the posterior is $\theta|Y, X \sim N(\tilde{\theta}, \tilde{\Sigma})$ with

$$\begin{aligned} \tilde{\theta} =& (X'X + \frac{I_k}{\tau^2})^{-1} X'Y \\ \tilde{\Sigma} =& (X'X + \frac{I_k}{\tau^2})^{-1} \end{aligned}$$

Inequality test: We want to test $H_0 : \theta < 0$ vs $H_1 : \theta \geq 0$. Assume $a_1 = a_2$, we would accept H_0 if:

$$\begin{aligned} P(\theta < 0|\mathcal{Y}_T) =& P\left(\frac{\theta - \tilde{\theta}}{\sqrt{\tilde{\Sigma}}} < -\frac{\tilde{\theta}}{\sqrt{\tilde{\Sigma}}}|\mathcal{Y}_T\right) && > \frac{1}{2} \\ =& \Phi\left(-\frac{\tilde{\theta}}{\sqrt{\tilde{\Sigma}}}\right) && > \frac{1}{2} \\ \tilde{\theta} <& && 0 \end{aligned}$$

With $\alpha = 5\%$, the frequentist test is based on:

$$\begin{aligned} \frac{\tilde{\theta}}{\sqrt{(X'X)^{-1}}} <& 1.64 \\ \tilde{\theta} <& 1.64\sqrt{(X'X)^{-1}} \end{aligned}$$

The Bayesian test of this hypothesis is consistent in the sense that as the sample size grows, we get the correct answer with probability 1. In a frequentist world, the probability of getting the right answer for the true null is called 1-size and approaches $1 - \alpha$.

Cite as: Anna Mikusheva, course materials for 14.384 Time Series Analysis, Fall 2007. MIT OpenCourseWare (http://ocw.mit.edu), Massachusetts Institute of Technology. Downloaded on [DD Month YYYY].

Point null: If we want to test $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$ and we use a continuous prior, then we will always reject the null. The solution is to specify a prior with a point mass of λ on 0. Take some continuous prior $p(\theta)$, let

$$p^*(\theta) = \lambda \Delta_{\theta=0} + (1-\delta)p(\theta)$$

where $\Delta_{\theta=0}$ is Dirac measure, then

$$\begin{aligned} p(\mathcal{Y}_T) =& \lambda f(\mathcal{Y}_T|0) + (1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta \\ p(\theta = 0|\mathcal{Y}_T) =& \frac{\lambda f(\mathcal{Y}_T|\theta=0)}{p(\mathcal{Y}_T)} \end{aligned}$$

The posterior odds ratio of the null to the alternative is

$$\frac{\lambda f(\mathcal{Y}_T|\theta=0)}{(1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta}$$

One difficulty here is in taking the integral in the denominator. Instead we can use the following trick: in the model without a mass point, we had $p(\mathcal{Y}_t) = \int f(\mathcal{Y}_T|\theta)p(\theta)d\theta = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\theta|\mathcal{Y}_T)}$.

Continue our OLS example . How consider their a test $H_0 : \theta = 0$ vs $H_1 : \theta \neq \theta_0$ with point mass $\lambda = 1/2$. We know that $f(\mathcal{Y}_T|\theta)$ is normal $N(\theta X_T, I)$, prior $p(\theta)$ is $N(0, \tau^2 I)$ and hence the posterior is also normal with mean and variance calculated before. We can plug it in $\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta = \frac{f(\mathcal{Y}_T|\theta)p(\theta)}{p(\theta|\mathcal{Y}_T)}$ and get $\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta$. After some calculation we can obtain that the posterior odds is equal to

$$\frac{\lambda f(\mathcal{Y}_T|\theta=0)}{(1-\lambda)\int f(\mathcal{Y}_T|\theta)p(\theta)d\theta} = \tau^k |X'X + \frac{I}{\tau^2}|^{1/2} \exp\left(-\frac{1}{2}\left(Y'X\left(X'X + +\frac{I}{\tau^2}\right)^{-1} X'Y\right)\right)$$

If costs $a_1 = a_2$ then the null is accepted if posterior odds are bigger than 1.

We will compare this test to the frequentist by studying the asymptotics of the log posterior odds ratio as $T \to \infty$. Assume that $\frac{X'X}{T} \to Q$

$$\log(po) = k \ln \tau + \frac{k}{2}\ln T + \frac{1}{2}\ln|\frac{X'X}{T} + \frac{I}{T\tau^2}| - \frac{1}{2}\left(Y'X\left(X'X + +\frac{I}{\tau^2}\right)^{-1} X'Y\right)$$

$k \ln \tau$ is a constant, and $\frac{1}{2}\ln|\frac{X'X}{T} + \frac{I}{T\tau^2}|$ converges to a constant so we may ignore them. We are interesting in whether $\lim \log(po) > 0$. Suppose $\theta_0 = 0$. Then,

$$\begin{aligned} Y'X\left(X'X + +\frac{I}{\tau^2}\right)X'Y =& \frac{Y'X}{\sqrt{T}}\left(\frac{X'X}{T} + \frac{I}{T\tau^2}\right)^{-1}\frac{X'Y}{\sqrt{T}} \\ \Rightarrow & N(0,\Sigma)Q^{-1}N(0,\Sigma)' \end{aligned}$$

This last expression is asymptotically bounded, so $\log(p.o.) = k \ln T + O_p(1) \to \infty$ and we accept the null asymptotically.

Now suppose $\theta_0 \neq 0$. Then $\frac{Y'X}{T} \to \theta_0 \frac{X'X}{T} \to \theta_0 Q$, so

$$\begin{aligned} Y'X\left(X'X + \frac{I}{\tau^2}\right)X'Y =& T\frac{Y'X}{T}\left(\frac{X'X}{T} + \frac{I}{T\tau^2}\right)\frac{X'Y}{T} \\ \to & T\theta_0 Q(Q)^{-1}Q'\theta_0' \\ \to & T\theta_0 Q\theta_0' \end{aligned}$$

so this term (which is negative in $\ln po$) asymptotically dominates $\log(p.o) = -T\theta_0 Q\theta_0'/2 + k\ln T + O_p(1) \to -\infty$and we reject the null asymptotically. Again, this test is asymptotically consistent, but frequentist tests are not. Consider what the frequentist test would be in this situation. The likelihood ratio is:

$$\begin{aligned} LR =& 2\ln \frac{f(Y|X,\hat{\theta}_{ML})}{f(Y|X,\theta=0)} \\ =& Y'X(X'X)^1X'Y \Rightarrow \chi^2 \end{aligned}$$

In particular, both Bayesian and frequentist test look at $Y'X(X'X)^1X'Y$ and reject if it is greater than some number, but Bayesians use $k\ln T$ and frequentists use χ^2 critical value. As a result, Bayes test is consistent, but frequentist test has a power against local alternatives $\theta = c/\sqrt{T}$ (while Bayes not).

To understand why the Bayesian test is consistent and the frequentist one is not, we need two asymptotic facts,

1. CLT: $\frac{1}{\sqrt{T}}\sum y_t \Rightarrow N(0,1)$

2. Law of iterated logarithm: $\frac{1}{\sqrt{2T\ln\ln T}}\sum y_t$ is almost surely asyptotically in $\in [-1,1]$, meaning that this sum does not converge, but for any number in $[-1,1]$, we can find a subsequence converging to that number, and the limits of all converging subsequences are in $[-1,1]$. In particular, almost surely (for any realization) there are infinitely many T such that $\frac{1}{\sqrt{2T\ln\ln T}}\sum_{t=1}^T y_t > 1-\delta$ for any δ, so we can always find infinitely many samples that will reject a null if we use frequentist a test.

   Cynical interpretation: every null can be rejected as long as enough data is collected:)