Bayesian Econometrics

Christopher A. Sims
Princeton University
sims@princeton.edu

September 20, 2016

Outline

I. The difference between Bayesian and non-Bayesian inference.

II. Confidence sets and confidence intervals. Why?

III. Bayesian interpretation of frequentist data analysis.

IV. A case where Bayesian and frequentist approaches seem aligned: SNLM

Bayesian Inference is a Way of Thinking, Not a Basket of "Methods"

- Frequentist inference makes only pre-sample probability assertions.
  - A 95% confidence interval contains the true parameter value with probability .95 only *before* one has seen the data. After the data has been seen, the probability is zero or one.
  - Yet confidence intervals are universally interpreted in practice as guides to *post*-sample uncertainty.
  - They often are reasonable guides, but only because they often are close to posterior probability intervals that would emerge from a Bayesian analysis.

- People want guides to uncertainty as an aid to decision-making. They want to characterize uncertainty *about parameter values*, given the sample that has actually been observed. That it aims to help with this is the distinguishing characteristic of Bayesian inference.

Abstract decision theory

Unknown state S taking values in Ω. We choose δ from a set Δ. Payoff is $U(\delta, S)$.

Admissible δ: There is no other $\delta' \in \Delta$ such that $U(\delta', S) > U(\delta, S)$ for all S.

Bayes decision rule δ^*: maximizes $E_P[U(\delta, S)]$ over δ for some probability distribution P over Ω.

The complete class theorem

- See Ferguson (1967)
- Under fairly general conditions, every admissible decision procedure is Bayes.
- This is, as mathematics, almost the same as the result that every efficient allocation in general equilibrium corresponds to a competitive equilibrium.
- Both are “separating hyperplanes” results, with the hyperplane defined by prices in general equilibrium and by probabilities in the complete class theorem.

A two-state graphical example

prior and model

Suppose we observe an animal dashing into a hole underneath a garden shed, and then note that there is no particularly bad odor around the hole. We are sure it is a woodchuck or a skunk.

Before we made these observations, our probability distribution on a randomly occuring animal in our back yard was .5 on woodchuck and .5 on skunk. This was our **prior** distribution.

Conditional on the animal being a woodchuck, we believe the probability that there would be odor around the burrow ($O = 1$) is .3, and the probability that the animal would run away fast ($R = 1$) is .7.

Conditional on the animal being a skunk, we believe $P[O = 1 \mid \text{skunk}] = .7$ and $P[R = 1 \mid \text{skunk}] = .3$

We think these observations are independent.

Tables of probabilities

Conditional probablities of observations, given woodchuck and given skunk.

	Woodchuck		Skunk	
	$O = 0$	$O = 1$	$O = 0$	$O = 1$
$R = 0$	.21	.09	.21	.49
$R = 1$	.49	.21	.09	.21

Joint distribution of parameters and data, posteior distribution

Therefore, the probability of what we saw is $P[O = 0 \text{ and } R = 1 \mid \text{woodchuck}] = .49$ and $P[O = 0 \text{ and } R = 1 \mid \text{skunk}] = .09$. To get the unconditional probabilities of these two events we multiply them by the prior probabilities of the parameters, which in this case is .5 for each of them.

Therefore the conditional probability of the animal being a woodchuck given our observations is .49/(.49+.09) =.845.

Confidence sets?

Is this also an 84.5% confidence set for the animal's species?

No, for two reasons. A confidence set is a random set that varies with the observed data. We can't determine whether this set $\{\text{woodchuck}\}$ is the value of a confidence set without specifying how the set would vary if we had made other observations. But besides that, there is no necessary connection between Bayesian posterior probabilities for intervals and confidence levels attached to them.

Constructing a confidence set

For example, we might specify that when we see $R = 0, O = 1$ our confidence set is just $\{$skunk$\}$, that when we see $R = 1, O = 0$ it is $\{$woodchuck$\}$, and that in the other two cases (where our posterior probabilities would be .5 on each) it is $\{$woodchuck,skunk$\}$. This would be a 90% confidence set (.49+.21+.21), though it would often leave us asserting with "90 per cent confidence" that the animal is either a woodchuck or skunk, even though we are quite sure that it is either one or the other.

Effect of a different prior

Our Bayesian calculations were based on the assumption that woodchucks and skunks are equally likely. If we knew before our observations that woodchucks are much more common than skunks, say 10 times more common, our marginal probability would put probability 10/11 on woodchuck. Then the probability of our $R = 1, O = 0$ observation is $.49 \cdot 10/11$ conditional on woodchuck, $.09 \cdot 1/11$ on skunk. Our observations would therefore move us from a 10/11=.909 probability on woodchuck to 4.9/(4.9+.09) = .982.

If instead we had seen $R = 0$, $O = 1$, the evidence would have shifted us from .909 probability on woodchuck to .049/(.049 + .09)=.353,

Model, parameters, prior, likelihood

This framework is not essential to Bayesian decision theory, but it is nearly universal in empirical work.

There are some unknown objects we label "parameters", in a vector β. There is a conditional distribution of observable data y given β, $p(y \mid \beta)$. The function $p(\cdot \mid \cdot)$ is the **model**. Frequentists and Bayesians agree on this setup. They also agree on calling $p(y \mid \beta)$, as a function of β with y fixed at the observed value, the **likelihood function**.

The formal difference: Bayesians treat y as non-random once it has been observed, but treat β as random (since it is unknown) both before and after y has been observed. Frequentists persist in making probability statements about y and functions of y (like estimators) even after y has been observed.

Bayes' Rule

With a pdf $\pi()$ describing uncertaintly about β before y has been observed (this is the **prior** pdf), The joint distribution of y, β has pdf $\pi(\beta)p(y \mid \beta)$. Bayes rule then uses this to calculate the conditional density of $\beta \mid y$ via

$$q(\beta \mid y) = \frac{\pi(\beta)p(y \mid \beta)}{\int \pi(\beta')p(y \mid \beta')d\beta'}.$$

Bayes' rule is math; not controversial. What bothers frequentists is the need for π and for treating the fixed object β as "random".

Is the difference that Bayesian methods are subjective?

- No.
- The objective aspect of Bayesian inference is the set of rules for transforming an initial distribution into an updated distribution conditional on observations. Bayesian thinking makes it clear that for decision-making, pre-sample beliefs are therefore in general important.
- But most of what econometricians do is not decision-making. It is reporting of data-analysis for an audience that is likely to have diverse initial beliefs. =
- In such a situation, as was pointed out long ago by Hildreth (1963) and Savage (1977, p.14-15), the task is to present useful information about the shape of the likelihood.

How to characterize the likelihood

- Present its maximum.
- Present a local approximation to it based on a second-order Taylor expansion of its log. (Standard MLE asymptotics.)
- Plot it, if the dimension is low.
- If the dimension is high, present slices of it, marginalizations of it, and implied expected values of functions of the parameter. The functions you choose might be chosen in the light of possible decision applications.

- The marginalization is often more useful if a simple, transparent prior is used to downweight regions of the parameter space that are widely agreed to be uninteresting.

HPD sets, marginal sets, frequentist problems with multivariate cases

Confidence set simple example number 1

- $X \sim U(\beta + 1, \beta - 1)$
- β known to lie in (0,2)
- 95% confidence interval for β is $(X - .95, X + .95)$
- Can truncate to (0,2) interval or not — it's still a 95% interval.

Example 1, cont.

- Bayesian with $U(0,2)$ prior has 95% posterior probability ("credibility") interval that is generally a subset of the intersection of the $X \pm 1$ interval with (0,2), with the subset 95% of the width of the interval.

- Frequentist intervals are simply the intersection of $X \pm .95$ with (0,2). They are wider than Bayesian intervals for X in $(0,2)$, but narrower for X values outside that range, and in fact simply vanish for X values outside $(-.95, 2.95)$.

Are narrow confidence intervals good or bad?

- A 95% confidence interval is always a collection of all points that fail to be rejected in a 5% significance level test.
- A completely empty confidence interval, as in example 1 above, is therefore sometimes interpreted as implying “rejection of the model”.
- As in example 1, an empty interval is approached as a continuous limit by very narrow intervals, yet very narrow intervals are usually interpreted as implying very precise inference.

Confidence set simple example number 2

- $X = \beta + \varepsilon$, $\varepsilon \sim e^{-\varepsilon}$ on $(0, \infty)$.
- likelihood: $e^{-X+\beta}$ for $\beta \in (-\infty, X)$
- Bayesian credible sets show reversed asymmetry
- contrast with naive (but commonly applied) bootstrap, which would produce an interval entirely concentrated on values of β *above* $\hat{\beta}_{MLE} = X$, which are impossible.
- (naive parametric bootstrap: Find an estimator $\hat{\beta}$. Simulate draws from $X \mid \hat{\beta}$, take the 2.5% tails of this distribution as the 95% confidence interval.)

Confidence set simple example number 3

- $Y \sim N(\beta, 1)$; we are interested in $g(\beta)$, where g is nonlinear and monotone, but unknown.
- We observe not Y, but $X = g(Y)$.
- We have a maximum likelihood estimator $\hat{g}$ for $g(\beta)$.
- If we could observe Y, a natural confidence and credible interval for β would be $Y \pm 1.96$. If we also knew g, we could then use $(g(Y - 1.96), g(Y + 1.96))$ as an excellent confidence and credible interval.

- Using the naive bootstrap here, if we base it on an estimator $\hat{g}$ asymptotically equivalent to the MLE, gives exactly the natural interval.

- So, without an analysis of likelihood, there is no answer as to whether the naive bootstrap gives reasonable results.

Mueller-Norets bettable confidence intervals

- If not bettable from either side, they're Bayesian for some prior.
- If only not bettable *against*, they exist, and are expanded versions of Bayesian posterior probability intervals.

Bayesian interpretation of frequentist asymptotics

- A common case: $\sqrt{T}(\hat{\beta}_T - \beta) \mid \theta \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \Sigma)$
- Under mild regularity conditions, this implies $\sqrt{T}(\beta - \hat{\beta}_T) \mid \hat{\beta}_T \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \Sigma)$
- Kwan (1998) presents this standard case (though his main theorem does not make strong enough assumptions to deliver his result). Kim (2002) extended the results beyond the $\sqrt{T}$ normalization case and thereby covered time series regression with unit roots.

SNLM

The SNLM often denoted by the equation $Y = X\beta + \varepsilon$, asserts the following conditional pdf for the vector of Y data conditional on the matrix of X data and on the parameters β, σ^2:

$$p(\underset{T\times 1}{Y} \mid \underset{T\times k}{X}) = \varphi(Y - X\beta; \sigma^2 I) = (2\pi)^{-T/2}\sigma^{-T}\exp\left(\frac{(Y - X\beta)'(Y - X\beta)}{2\sigma^2}\right)$$

The most common framework for Bayesian analysis of this model asserts a prior that is flat in β and $\log\sigma$ or $\log\sigma^2$, i.e. $d\sigma/\sigma$ or $d\sigma^2/\sigma^2$. We will assume the prior has the form $d\sigma/\sigma^p$, then discuss how the results depend on p.

Marginal for σ^2

We let $u(\beta) = Y - X\beta$ and denote the least squares estimate as $\hat{\beta} = (X'X)^{-1}X'Y$. Also $\hat{u} = u(\hat{\beta})$. Then the posterior can be written, by multiplying the likelihood above by σ^{-p} and rearranging, as proportional to

$$\sigma^{-T-p} \exp\left(-\frac{\hat{u}'\hat{u} + (\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{2\sigma^2}\right) d\beta\, d\sigma$$

$$\propto \sigma^{-T-p+k} |X'X|^{-\frac{1}{2}} \exp\left(-\frac{\hat{u}'\hat{u}}{2\sigma^2}\right) \varphi\left(\beta-\hat{\beta}; \sigma^2(X'X)^{-1}\right) d\beta\, d\sigma$$

$$\propto v^{(T+p-k)/2} \exp\left(\frac{-\hat{u}'\hat{u}}{2}v\right) |X'X|^{-\frac{1}{2}} \varphi\left(\beta-\hat{\beta}; \sigma^2(X'X)^{-1}\right) d\beta \frac{dv}{v^{3/2}},$$

where $v = 1/\sigma^2$.

Integrating this expression w.r.t. β and setting $\alpha = \hat{u}'\hat{u}/2$ gives us an expression proportional to

$$v^{(T+p-k-3)/2} \exp\left(-\frac{\hat{u}'\hat{u}}{2}v\right) dv \propto \alpha^{(T+p-k-1)/2} v^{(T+p-k-3)/2} e^{-\alpha v} dv\,,$$

which is a standard $\Gamma((T+p-k-1)/2, \alpha)$ pdf.

Because it is $v = 1/\sigma^2$ that has the Γ distribution, we say that σ^2 itself has an **inverse-gamma** distribution. Since a $\Gamma(n/2, 1)$ variable, multiplied by 2, is a $\chi^2(n)$ random variable, some prefer to say that $\hat{u}'\hat{u}/\sigma^2$ has a $\chi^2(T-k)$ distribution, and thus that σ^2 has an inverse-chi-squared distribution.

Marginal on β

Start with the same rearrangement of the likelihood in terms of v and β, and rewrite it as

$$v^{(T+p-3)/2} \exp\left(-\frac{1}{2}u(\beta)'u(\beta)v\right) dv\, d\beta\,.$$

As a function of v, this is proportional to a standard $\Gamma\big((T+p-1)/2, u(\beta)'u(\beta)/2\big)$ pdf, but here there is a missing normalization factor that depends on β. When we integrate with respect to v, therefore, we arrive at

$$\left(\frac{u(\beta)'u(\beta)}{2}\right)^{-(T+p-1)/2} d\beta \quad \propto \quad \left(1+\frac{(\beta-\hat{\beta})'X'X(\beta-\hat{\beta})}{\hat{u}'\hat{u}}\right)^{-(T+p-1)/2} d\beta\,.$$

This is proportional to what is known as a multivariate $t_n\big(0, (\hat{u}'\hat{u})/n\big)$ pdf, where $n = T + p - k - 1$ is the degrees of freedom. It makes each element β_i of β an ordinary univariate $t_n(\hat{\beta}_i, s_\beta^2)$, where $s_\beta^2 = s^2(X'X)_{ii}^{-1}$ and $s^2 = \hat{u}'\hat{u}/n$. Thus the statistics computed from the data can be analyzed with the same tables of distributions from either a Bayesian or non-Bayesian perspective.

Breaks

- Suppose we have a model specifying Y_t is i.i.d. with a pdf $f(y_t; \mu_t)$, and with μ_t taking on just two values, $\mu_t = \underline{\mu}$ for $t < t^*$, $\mu_t = \bar{\mu}$ for $t \geq t^*$.

- The pdf of the full sample $\{Y_1, \ldots, Y_T\}$ therefore depends on the three parameters, $\underline{\mu}, \bar{\mu}, t^*$.

- If f has a form that is easily integrated over μ_t and we choose a prior $\pi(\underline{\mu}, \bar{\mu})$ that is conjugate to f (meaning it has the same form as the likelihood), then the posterior marginal pdf for t^* under a flat prior on t^* is easy to calculate: for each possible value of t^*, integrate the posterior over $\underline{\mu}, \bar{\mu}$.

- The plot of this integrated likelihood as a function of t^* gives an easily interpreted characterization of uncertainty about the break date.

- Frequentist inference about the break date has to be based on asymptotic theory, and has no intepretation for any observed complications (like multiple local peaks, or narrow peaks) in the global shape of the likelihood.

Model comparison

- Can be thought of as just estimating a discrete parameter: $y \sim f(y \mid \theta, m)$, with $\theta \in \mathbb{R}^k$, $m \in \mathbb{Z}$.
- Then apply Bayes rule and marginalization to get posterior on m:

$$p(m \mid y) \propto \int f(y \mid \theta, m)\, d\theta\,.$$

- The rhs is the **Bayes factor** for the model.
- This is the right way to compare models in principle, but it can be misleading if the discrete parameter arises as an approximation to an underlying continuous range of uncertainty — as we'll discuss later.

Simple examples of model comparison, vs. "testing" models

- Test $H0 : X \sim N(0,1)$. "reject" if $|X|$ too big.
- vs. what? $HA : X \sim N(0,2)$? $N(0,.5)$?
- The classic frequentist testing literature emphasized that tests must always be constructed with the alternative hypothesis in mind. There is no such thing as a meaningful test that requires no specification of an alternative.
- Posterior odds vs. 5% test for $N(0,1)$ vs. $N(0,2)$.
- $N(0,1)$ vs. $N(2,1)$, $N(\mu,1)$? Need for prior.

The stopping rule paradox

- Boy preference birth ratio example
- Sampling to significance example
- Posterior distribution vs. test outcome

*

References

FERGUSON, T. S. (1967): *Mathematical Statistics: A Decision Theoretic Approach*. Academic Press, New York and London.

HILDRETH, C. (1963): "Bayesian Statisticians and Remote Clients," *Econometrica*, 31(3), 422–438.

KIM, J.-Y. (2002): "Limited information likelihood and Bayesian analysis," *Journal of Econometrics*, 107, 175–193.

KWAN, Y. K. (1998): "Asymptotic Bayesian analysis based on a limited information estimator," *Journal of Econometrics*, 88, 99–121.

SAVAGE, L. J. (1977): "The Shifting Foundations of Statistics," in *Logic, Laws and Life*, ed. by R. Colodny, pp. 3–18. University of Pittsburgh Press.