Bayesian Econometrics

Christopher A. Sims
Princeton University
sims@princeton.edu

November 5, 2019

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 $\delta : \Omega \to Q$ from a set Δ. Payoff is $U(\delta(S), S)$.

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

Bayes decision rule δ^*: maximizes $E_P[U(\delta(S), S)]$ 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.

Reminder: General form of Bayesian inference

- Collection of random variables divided into two pieces: Y, θ.
- Joint density over Y, θ specified in the form of a conditional density for $Y \mid \theta$, $p(y \mid \theta)$, and a marginal density $\pi(\theta)$ for θ.
- $p(\ \mid\ )$ is called the model; $\pi(\ )$ is called the prior; $p(y \mid\ )$ as a function of θ with y fixed at the observed value is the likelihood.
- Y will be observed, θ will not be.
- Bayes' rule is applied to convert π into a conditional density for $\theta \mid Y$

$$\frac{p(y \mid \theta)\pi(\theta)}{\int p(y \mid \theta)\pi(\theta)\, d\theta}$$

Aside: Why separate model and prior?

- We're so used to the idea of a model that this may seem uncontroversial, but from a pure subjective decision-making perspective, the separation is unnecessary.

- We could just specify a joint distribution for Y, θ and proceed.

- The main justification for the separation is scientific communication: we may hope that the model is widely accepted as an accurate description of conditional uncertainty about Y (a "data-generating process"), while our audience may have differing views about a reasonable prior.

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

- In Bayesian inference, a useful notion is the **highest posterior density**, or HPD posterior probability set (sometimes called a credibility set)

- A 90% HPD set, for example, is the smallest set in the parameter space with posterior probability .9. All points inside it have values of the posterior pdf at least as large as for any point outside it.

- When such a set turns out to be small, it implies that there is little posterior uncertainty about the parameter.

- Another common type of posterior probability set: equal-tailed. A 90% equal-tailed credible set is an interval (not a more general set) with 5% probability above it and 5% below it.

Marginal probability sets

- If there are n parameters, but we are interested in $k < n$ of them, it is straightforward to construct an HPD set for the k parameters.

- One integrates the other parameters out of the posterior distribution, then uses the marginal distribution of the k parameters to construct an HPD set.

- There is no corresponding direct way to construct frequentist confidence regions for subsets of parameters.

- The standard normal linear model (SNLM) that underlies linear regression is an example where confidence sets for subsets of parameters are possible. But it is a knife-edge special case.

Difficulties for confidence sets

- The problem: A $100 \cdot (1 - \alpha)$ per cent confidence set must have a probability at least $1 - \alpha$ of containing the true parameter value, no matter what the true parameter value is. Generally this probability (called the **coverage probability**) varies with all the parameters, not just the k we are interested in.

- The result is that true confidence sets for subsets of parameters are often misleadingly big — because a few values of the remaining $n - k$ parameters make inference for the k we're interested in difficult, the confidence set has to be very wide.

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. Flat-prior Bayesian $1 - \alpha$ credible sets match frequentist likelihood-ratio based $1 - \alpha$ confidence sets.

Bootstrap CI"s?

- Suppose we have a sample of N i.i.d. draws of X. The MLE for β is the minimum of the observed X's.

- Pure bootstrap: Draw randomly with replacement from the observed values of X, creating a collection of size-N samples. For each such sample i, set $\hat{\beta}_i$ to the minimum of the X's in that sample. Take as CI for β the $\alpha/2$ and $(1-\alpha/2)$ tails of the empirical distribution of the $\hat{\beta}_i$ draws.

- Naive parametric bootstrap: Find an estimator $\hat{\beta}$. Simulate draws from $X \mid \hat{\beta}$, form interval as before.

- Both of these bootstraps produce a “CI” that is bounded below. The pure bootstrap or parameteric bootstrap with the MLE of β produce intervals that lie entirely at or above the observed minimum X, whereas the truth is certainly below this value.

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

- Suppose we propose a 90% confidence set — a mapping from realizations of the data to subsets of the parameter space — and a shrewd gambler gets to, for every observed sample, decide whether to bet, at 9 to 1 odds, that the true parameter is not in the confidence set.

- If the shrewd gambler has positive expected winnings, we might be dissatisfied with the confidence set.

- An obvious simple example: If the confidence set is empty with positive probability, the gambler can bet against just the empty confidence sets and make money.

If not bettable, CI's are either Bayesian for some prior, or contain a Bayesian credible set for some prior

- If not bettable from either side (gambler can choose to bet parameter is in or out), they're Bayesian for some prior.
- If only not bettable *against*, they exist, and are expanded versions of Bayesian posterior probability intervals — there is a prior such that all the 90% confidence intervals contain 90% credible sets for that prior.
- Note that our example 1 above is not bet-proof, as it can produce empty confidence sets.

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. Kim (2002) extended the results beyond the $\sqrt{T}$ normalization case and thereby covered time series regression with unit roots.

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: $y \sim p(y \mid \theta, m)$, with $\theta \in \mathbb{R}^k$, $m \in \mathbb{Z}$, prior $\pi(\theta, m)$.
- Then apply Bayes rule and marginalization to get posterior on m:

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

- 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.