Bayesian Methods in Applied Econometrics, or, Why Econometrics Should Always and Everywhere Be Bayesian

Christopher A. Sims
Princeton University
sims@princeton.edu

May 22, 2008

Introduction

- The first part of the talk makes some unoriginal claims about the role of Bayesian thinking.
- Despite being unoriginal, and obvious to me and to a minority of econometricians, they are unfamiliar, thought-provoking, or outrageous to quite a few econometricians.
- But it is worthwhile from time to time to restate claims that are both obvious and outrageous and that therefore (for one of these reasons or the other) are usually excluded from journal articles.
- The latter part of the talk discusses some areas of econometric application where frequentist asymptotics seems particularly persistent and suggests how Bayesian approaches might become more practical and prevalent.

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

1.1 What it is

Probability statements conditioned on observations

- 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.

Bayesian inference should be the starting point

- For discussing implications of data analysis with decision makers.
- They often are quite familiar with the language of probability and may ask, for example, for what the data say about the odds of a parameter being in one region vs. being in another.
- A carefully trained frequentist econometrician who knows only how to construct confidence regions and hypothesis tests must be tongue-tied in the face of such a request.
- The Bayesian approach to inference should be the starting point also of our education of econometricians. For the time being, they need also to learn what a confidence region is (what it *really* is, as opposed to what most of them think

it is after a one-year statistics or econometrics course). But I think that full understanding of what confidence regions and hypothesis tests actually are will lead to declining interest in constructing them.

1.2 Objections

It's subjective, whereas frequentist approaches are objective

- This is simply untrue.
- 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.

It requires stronger assumptions than frequentist asymptotics

- Bayesian small-sample inference based on specific parametric assumptions is of course approximately correct also when the parametric assumptions are approximately correct.
- Likelihoods often become approximately Gaussian in large samples, and distributions of estimators $\hat{\beta}$ then tend to become symmetric functions of $\hat{\beta} - \beta$, so that the likelihood and the pdf of the parameter have the same shape.
- The conditions that allow these claims about asymptotic likelihood shapes are almost the same as the "weak" conditions allowing asymptotic distribution claims for estimators.
- So in standard cases Bayesian inference is approximately correct in large samples under assumptions roughly as "weak" as those that make frequentist inference approximately correct in large samples.

Assumptions, continued: Asymptotic results under weak assumptions do not justify small-sample assertions under weak assumptions

- Consider kernel estimation of a regression function: $y_i = f(x_i) + \varepsilon_i$
- It will be consistent and allow accurate confidence statements for a large class of functions f — in large samples under weak assumptions.
- But in a particular sample, we have to pick a bandwidth. If f shows rapid changes over intervals shorter than the kernel bandwidth, our kernel estimator will be badly biased.
- The needed restriction on variability of f is not a "weak" assumption, and in frequentist inference it is not explicit.

Bayesian inference is hard

- I think it's Don Berry who said, "Bayesian inference is hard in the sense that thinking is hard."

- But it is true that Bayesian inference is usually described and often implemented in a bottom-up way: define your model, define your prior, apply Bayes rule, emerge with unique, optimal, correct inference.
- This is indeed often hard. Not just computationally, in the third step, but intellectually, in the first two.
- Frequentist inference could be approached in the same way: Define your model, derive fully efficient estimators, pay no attention to anything else.
- This is actually even harder, and furthermore is not guaranteed to give you any answer, or a unique answer as to how to proceed.

Hard, continued: Frequentists take shortcuts because doing it right is *harder* for them.

- It is standard practice for frequentists to write papers about methods that are convenient, or intuitively appealing, and describe their properties, often only approximately, based on asymptotics.
- People use such procedures, often in preference to more elaborate fully efficient procedures.
- Bayesians can also analyze the properties of estimators or statistics that are not optimal given the model and prior.
- They also can use asymptotic approximation as a shortcut when exact derivations are hard or as a way of avoiding the need to commit fully to all the details of a probability model.
- They should probably do more of this, and also stop claiming that it is an advantage of Bayesian procedures that they give "exact small sample results".

Frequentists can test hypotheses without specifying the alternative

- Neymann would not have agreed.
- There is always at least an implicit alternative, and there are always alternatives under which a given test would be senseless.

- Nonetheless hypothesis testing is often taught without attention to dependence of the test's usefulness on the alternative.
- The view that formal econometrics leads to "testing" and "rejecting" models without presenting an alternative is part of what has given econometrics a bad name in some quarters (e.g. among macro calibraters).

One-handed tests, continued

- So it is an advantage of Bayesian approaches that they make it clear that one can only emerge from data analysis with odds ratios of models against one another, not with a "test" of a model in isolation.
- (Some Bayesians, in the area of "Bayesian model validation", come perilously close to trying to produce alternative-free "tests" with Bayesian machinery.)

1.3 Frequentist methods from a Bayesian perspective

Frequentist asymptotics usually imply Bayesian asymptotics

- If $\sqrt{T}(\hat{\beta} - \beta) \mid \beta \xrightarrow{D} N(0, \Sigma)$, with Σ not dependent on β, this usually implies that $\sqrt{T}(\beta - \hat{\beta}) \mid \hat{\beta} \xrightarrow{D} N(0, \Sigma)$.
- That is, frequentist approximate confidence statements based on asymptotics are usually interpretable as Bayesian probability statements about parameter location *conditional on the estimator* — rather than conditional on the full sample.
- Such asymptotic results, in other words, can be interpreted as delivering limited information approximate Bayesian inference. (See Kwan (1998) and Kim (2002).)

But much of frequentist econometric theory looks like wasted effort

- For example, "testing" of compound hypotheses where probabilities of rejection vary across the null set.
- Or most of the unit root literature.

- Or most of the "testing for breaks" literature. (In the simple case of a single break for a regression model, one can easily compute integrated posteriors for each possible break date, and the plot of these gives the posterior for the break date under a uniform prior. No analysis of the asymptotic behavior of the likelihood as a stochastic process on the break dates conditional on the break date is needed.)
- These are instances where tracing out the implications of the likelihood is quite straightforward in a given sample, but frequentist results are much harder — and less useful.

Good frequentist practice has a Bayesian interpretation

- This is a definition, not a theorem.
- Working out the priors and modeling assumptions that would justify a frequentist procedure as Bayesian, or approximately Bayesian, can be helpful.
- It can reassure us (at least us Bayesians) that the procedure makes sense.
- It can turn up situations in which an implied prior has strange characteristics, and thereby lead us to better estimates or procedures.
- It can help us identify rare conditions or samples in which the frequentist procedure deviates sharply from its Bayesian approximant, and thereby again lead us to better estimates or procedures.

2 Recent Successes

Macro policy modeling

- The ECB and the New York Fed have research groups working on models using a Bayesian approach and meant to integrate into the regular policy cycle. Other central banks and academic researchers are also working on models using these methods.
- Bayesian methods have gained increasing attention in this area because:
  - the modeling is directly tied to repeated decision-making use;

- it involves large models with many parameters, so that attempting to proceed without any use of prior distributions is a dead end;
- computational power and MCMC methods now make Bayesian analysis of large models feasible.

What is MCMC?

- This is not the place for a detailed explanation.
- It is a method of posterior simulation.
- Its appeal is that whenever the posterior density function can be computed for arbitrary points in the parameter space, it is possible with MCMC to generate simulated samples from that posterior density, even though the density corresponds to no known distribution.

Mixed models

In statistics, **mixed models**, estimated by MCMC, have received a lot of attention. They should receive more attention from economists.

$$y_{it} = X_{it}\beta_i + u_{it}$$
$$\beta_i \sim N(\bar{\beta}, \Omega)$$
$$u_{it} \sim N(0, \sigma^2)$$

This type of model is easy to handle by MCMC, harder to make sense of from a frequentist perspective because it has "parameters" that are "random". It should probably be preferred in most applications to trying to control for conditional heteroscedasticity via "cluster" corrections on standard errors.

3 Frontiers: "Weak Assumptions"

Areas of econometrics where Baysian approaches are rare

- In cross-section and panel data econometrics frequentist theory and practice remain dominant.
- Instrumental variables, GMM, and non-parametric modeling are widely used, and there is a general impression that Bayesians have no substitute for them.

3.1 IV,GMM

IV

- IV estimates and confidence bands generally satisfy the regularity conditions allowing their interpretation as approximate Bayes estimators and probability intervals.
- The sample moments from which IV estimators are constructed are sufficient statistics under the model that leads to the LIML likelihood.
- A natural Bayesian approach, then, which improves on the usual IV procedures, is to characterize the shape of the LIML likelihood, perhaps using a prior that captures consensus views on what are interesting parameter values.
- This approach handles "weak instruments" transparently. Samples with a weak instruments problem are samples where the posterior density contours are far from the usual Gaussian eggs in regions of the parameter space that are a priori interesting.
- When there are large numbers of instruments, a prior will help in making sense of results, and in avoiding the trivial result from 2SLS when the first stage has negative or small degrees of freedom.
- See my paper "Thinking about Instrumental Variables", available on my website.

But doesn't this involve making strong distributional assumptions?

- The trivial answer just points to the fact that IV, LIML, and Bayesian posterior means all will agree asymptotically.
- So the Bayesian posterior probability intervals calculated with "strong assumptions" have the same asymptotic validity, under the same "weak" assumptions, as the asymptotic-theory based intervals for IV.
- The Bayesian analysis makes clear exactly what assumptions are needed to make "asymptotically justified" probability statements actually justified.

Conservative models

- But we can go farther. We can ask, if given the moment assumptions on which IV is based, what is the most "conservative" small-sample probability model satisfying those assumptions?
- A natural formulation: minimize the *mutual information*, in Shannon's sense, between the model parameters and the data.
- The LIML Gaussian setup emerges from this kind of reasoning.
-

GMM

- Here a model and prior that would lead to the usual procedures is less apparent.
- Suppose our moment conditions are $E[g(y;\beta) \mid \beta] = 0$.
- Usually we start with $E[g_0(\gamma) \mid \gamma, \Sigma)] = 0$, and add

$$g_1(y;\gamma,\Sigma) = g_0(y;\gamma,\Sigma)g_0(y;\gamma,\Sigma)' - \Sigma$$

  .
- Our g includes both g_0 and g_1 and our β includes both γ and Σ.

Mutual information minimization

With $p(y \mid \beta)$ the pdf of i.i.d. y's and π the prior pdf on β, minimize w.r.t. $p(y \mid \beta)$

$$-\int \log\left(\int p(y \mid \beta')\pi(\beta')d\beta'\right) p(y \mid \beta)\pi(\beta)d\beta\, dy \\ + \int log(p(y \mid \beta))p(y \mid \beta)dy\, \pi(\beta)d\beta \quad (1)$$

subject to, for each β,

$$\int p(y \mid \beta)dy = 1 \qquad \int g(y \mid \beta)p(y \mid \beta)dy = 0\,.$$

Form of the joint pdf

We emerge with the joint pdf

$$q(y, \beta) = e^{A(\beta)+B(y)+C(\beta)g(y|\beta)} .$$

- This implies that the posterior depends on the data only through $\sum g(y_i \mid \beta)$.
- The linear IV setup, leading to normality, is a special case.
- The result generally depends, through A, B, and C, on the prior.
- It does not, when $\partial g / \partial y$ is non-constant, generally lead to a model in which g_0 is $N(0, \Sigma)$.
- If one can find a $B(y)$ and $C(\beta)$ that makes the integral $p(\beta) = \int \exp\big(C(\beta)g(y \mid \beta) + B(y)\big)dy$ finite for each β and at the same time makes $E[g(y \mid \beta) \mid \beta)] = 0$ for each β under the conditional pdf for y derived from the joint density, the problem is solved. But finding such a B, C pair for general nonlinear g is not easy.

Form of the joint pdf, cont.

- What about just using the GMM objective function, scaled appropriately by $|\Sigma|$, i.e. $|\Sigma|^{-T/2} \exp\big(-(\sum g_0)'\Sigma^{-1}(\sum g_0)\big)$, as if it were a likelihood?
- The problem is that in general this implies a small sample distribution in which the moment conditions defining the GMM estimator are *not* satisfied.
- It seems worthwhile to explore in more detail what emerges from this setup.

3.2 Nonparametrics

What applied modelers do in practice

1. Specify a reasonable-looking model.
2. Estimate its parameters.
3. Apply "specification tests", by comparing the model to more complicated ones
4. If the specification tests so indicate, expand the model and go to (2).

Bayesian sieve

- Put a prior on an ∞-dimensional parameter space by putting probability one on a countable union of finite-dimensional parameter spaces.
- E.g., $B(L)y_t = \varepsilon_t$, $B(L)$ is of order n with probability 2^{-n}.
- This leads to doing essentially what was described on the previous slide — estimating finite-dimensional models, expanding or shrinking the number of parameters based on evidence of the extent to which larger models fit better.
- If the finite-dimensional spaces are dense in the full space, then under some regularity conditions we can get consistency in the full space.

A countable union of finite-d subspaces is a small piece of an ∞-d topological vector space

- ∞-d TVS's are not locally compact.
- Finite-d TVS's are locally compact.
- Finite-d subspaces are nowhere dense.
- Countable unions of them are "meagre". One topological definition of "small".
- So this looks restrictive. We've put probability one on a small subset of the parameter space.

Every prior on an ∞-d space does this

- Probability measures on complete, separable metric spaces are "tight", meaning they put probability one on countable unions of compact sets.
- But this means they put probability one on subsets that are topologically small in the same sense that a countable union of finite-d subspaces is small.
- In an AR, where the natural topology of fit on the coefficients is equivalent to ℓ_2, another way to get a dense countable union of compact sets of $B(L)$'s is to require that for $|b_i| < A(B)i^{-q}$ for some fixed $q \geq 1$, for example.

But this is not a problem peculiar to Bayesian inference

- Asymptotically valid confidence statements require exactly the same kinds of restrictive assumptions.
- Kernel estimates impose bounds on tail behavior and/or derivatives.
- Conclusion 1: If you want to estimate an infinite-dimensional parameter from a countable sequence of finite-dimensional observations, you have to take a strong prior position, so you might be wrong forever in your probability assertions.
- Conclusion 2: Don't be embarrassed to proceed as usual with finite parameterizations, tests for specification error. Fancy non-parametric methods are not in fact any more general.

Bayesian "kernel" regression

$$y_t = f(x_t) + \varepsilon_t$$

ε_t i.i.d., zero mean conditional on $\{x_s\}$. The parameter is f. Assume f is a stochastic process on x as index set, e.g. Gaussian with $\text{Cov}(f(x), f(z)) = R_f(x - z)$. Then we have a covariance matrix for a sample of y's, a cross-covariance of the y's with the $f(x)$'s, and thus a projection formula for $E[f(x^*) \mid y_1, \ldots, y_T]$.

For x's in the midst of observed x_i's, the weights in the projection look very much like kernel weights. For x's at the boundary of the sample, or in areas where x_i's are very sparse, the kernel spreads out, becomes asymmetric. This corresponds to what applied workers actually do.

The plots on the next slide show the weights from an example in which the $R_f(x) = 1 - |x|\,k$ on $[-k, k]$, with $k = .12$. More detail is in Sims (2000), from which these graphs were extracted. One of the plots shows a case where the x^* value has many observed x_i near-neighbors. The other shows a case of an x^* at the boundary of the x range and without many near neighbors.

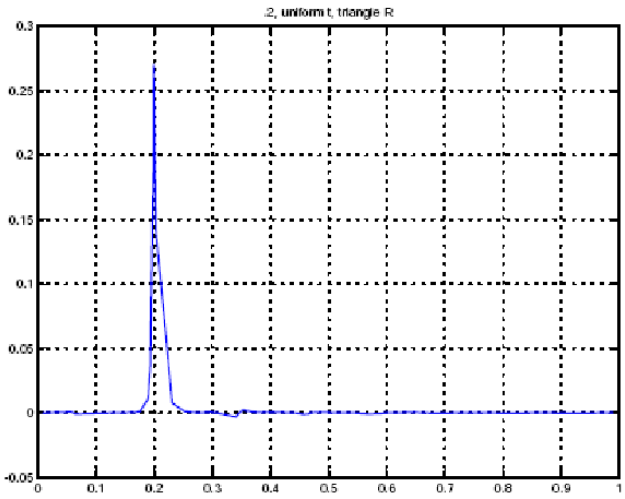

FIGURE 4. Weighting function for $x^* = .2$
Nearest x_i: 6 between .1934 and 0.2028

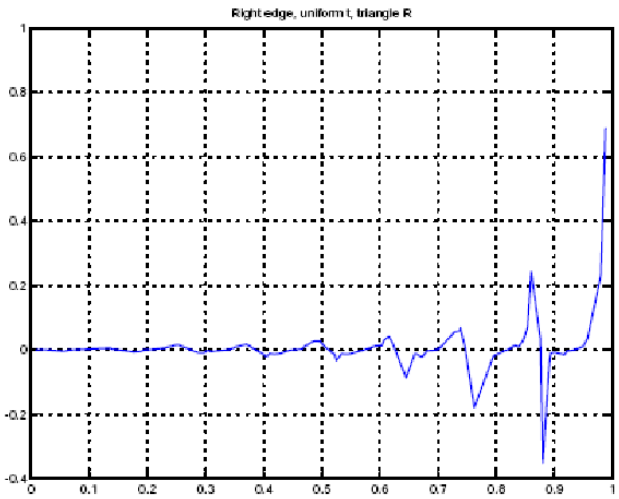

FIGURE 5. Weighting function for $x^* = 1$
Nearest x_i: .9501 .9568 .9797 .9883

Conclusion

Lose your inhibitions: Put probabilities on parameters without embarrassment.

References

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

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

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

SIMS, C. A. (2000): "Using a Likelihood Perspective to Sharpen Econometric Discourse: Three Examples" *Journal of Econometrics* 2000, 95(2), 443–462, `http://www.princeton.edu/~sims/`.