UNDERSTANDING NON-BAYESIANS

ABSTRACT.

I. INTRODUCTION

Once one becomes used to thinking about inference from a Bayesian perspective, it becomes difficult to understand why many econometricians are uncomfortable with that way of thinking. But some very good econometricians are either firmly non-Bayesian or (more commonly these days) think of Bayesian approaches as a "tool" which might sometimes be appropriate, sometimes not. This paper tries to articulate the counterarguments to a Bayesian perspective. There are some counterarguments that are frequently expressed, but are not hard to dismiss. Others, though, correspond to types of application where convenient, seemingly sensible, frequentist tools exist, while Bayesian approaches are either not yet developed or seem quite inconvenient. And there are also counterarguments that relate to deep questions about inference on infinite-dimensional parameter spaces and to corresponding pitfalls in the application of Bayesian ideas. Section II explains the difference between Bayesian and frequentist approaches to inference. Section III discusses commonly heard, but weak, objections, while section IV takes up subtler issues. Section V illustrates the subtler issues in the context of some specific models.

To relieve the reader's possible suspense: My conclusion is that the Bayesian perspective is indeed universally applicable, but that "non-parametric" inference is hard, in ways about which both Bayesians and non-Bayesians are sometimes careless.

II. THE DIFFERENCE BETWEEN BAYESIAN AND FREQUENTIST APPROACHES TO INFERENCE

Frequentist inference insists on a sharp distinction between unobserved, but non-random "parameters" and observable, random, data. It works entirely with probability distributions of data, conditional on unknown parameters. It considers the random behavior of functions of the data — estimators and test statistics, for example — and makes assertions about the distributions of those functions of the data, conditional on parameters.

Date: September 1, 2010.

Bayesian inference treats everything as random before it is observed, and everything observed as, once observed, no longer random. It aims at assisting in constructing probability statements about anything as yet unobserved (including "parameters") conditional on the observed data. Bayesian inference aids in making assertions like, "given the observed data, the probability that β is between one and two is .95", for example. Naive users of statistical methods sometimes think that this kind of assertion is what a frequentist confidence interval provides, but this is not true. For a careful frequentist, a confidence interval, once computed from a particular sample, either contains β or does not contain β. Once a particular sample has been observed, there is no random variation left to put probabilities on in a frequentist analysis.

Bayesian inference therefore feeds naturally into discussion of decisions that must be made under uncertainty, while frequentist analysis does not. There are theorems showing that under certain assumptions it is optimal to make decisions based on probability distributions over unknown quantities, but it is probably more important that most actual decision makers find the language of probability a natural way to discuss uncertainty and to consider how the results of data analysis may be relevant to their decisions. If an unknown β matters importantly to a decision, the decision maker is likely to be assessing probabilities that β lies in various regions, and is likely to be interested in what data analysis implies about those probabilities.

Careful frequentists understand these distinctions and sometimes correct careless characterizations of the meaning of confidence intervals, test, and estimator properties. Nonetheless the ease of convenient misinterpretation may be why frequentist inference focuses on functions of the data that are easily, and commonly, interpreted as if they were post-sample probability statements about the location of parameters. "Rejection of $H_0 : \beta = 0$ at the 5% level" is likely to be interpreted as "the data indicate that with 95% probability that β is nonzero". "$\hat{\beta}$ is an unbiased estimate of β" is likely to be interpreted as "the data indicate that the expected value of β is $\hat{\beta}$", and so on.

There are some widely applied models, like the standard normal linear regression model, in which Bayesian probability statements about parameters given the data have the same form, or almost the same form, as frequentist presample probability statements about estimators. In many more models Bayesian probability statements about parameters are nearly the same, in large samples, with high probability, as frequentist presample probability statements about estimators. Students who have had only a smattering of statistics therefore often form the mistaken impression that there is no important distinction between frequentist approaches to inference and

Bayesian approaches with "flat priors". The distinguishing feature of Bayesian inference then appears to be that Bayesians give explicit attention to non-flat priors that are based on subjective prior beliefs.

III. Some easily dismissed objections

III.1. **Bayesian inference is subjective.** Probably the most common mistaken characterization of the difference between Bayesian and frequentist inference locates the difference in subjectivity vs. objectivity: It is true that Bayesian inference makes the role of subjective prior beliefs in decision-making explicit, and describes clearly how such beliefs should be modified in the light of observations. But most scientific work with data leads to publication, not directly to decision-making. That is, most data analysis is aimed at an audience who face differing decision problems and may have diverse prior beliefs. In this situation, as was pointed out long ago by Hildreth (1963) and (Savage, 1977, p.14-15), useful data analysis summarizes the shape of the likelihood. Sometimes it is helpful to apply non-flat, simple, standardized priors in reporting likelihood shape, but these are chosen not to reflect the investigator's personal beliefs, but to make the likelihood description more useful to a diverse audience. A Bayesian perspective makes the entire shape of the likelihood in any sample directly interpretable, whereas a frequentist perspective has to focus on the large-sample behavior of the likelihood near its peak.

Though frequentist data analysis makes no explicit use of prior information, good applied work does use prior beliefs informally even if it is not explicitly Bayesian. Models are experimented with, and versions that allow reasonable interpretations of the estimated parameter values are favored. Lag lengths in dynamic models are experimented with, and shorter lag lengths are favored if longer ones add little explanatory power. These are reasonable ways to behave, but they are not "objective".

In short, researchers who take a Bayesian perspective can take a completely "objective" approach, by aiming at description of the likelihood. Frequentists have no formal interpretation of the global likelihood shape. Frequentist textbook descriptions of methods make no reference to subjective prior beliefs, but everyone recognizes that good applied statistical practice, even for frequentists, entails informal use of prior beliefs when an actual decision is involved. Its supposed "subjectivity" is therefore no reason to forswear the Bayesian approach to inference.

III.2. **Bayesian inference is harder.** There is nothing inherent in the frequentist perspective that implies its adherents must propose convenient or intuitively appealing

estimators and derive their asymptotic properties; it could instead insist on working only with the most efficient estimators available and working with exact, small-sample distribution theory. But, perhaps because verifiably maximally efficient frequentist estimators are often not available, or because small-sample distribution theory is dependent on parameter values in complicated ways, the majority of the frequentist literature does in fact derive asymptotic properties of convenient or intuitively appealing estimators.

It is generally easier to characterize optimal small-sample inference from a Bayesian perspective, and much of the Bayesian literature has insisted that this is a special advantage of the Bayesian approach. Furthermore, the recent spread of conceptually simple, computer-intensive methods of simulating large samples from posterior distributions, like Markov chain Monte Carlo methods, has made characterization of likelihood shape relatively straightforward even in large non-linear models.

Nonetheless there is no reason in principle that very easily implemented estimators like instrumental variables estimators or GMM estimators or kernel-smoothed regression estimators have to be interpreted from a frequentist perspective. Kwan (1998) showed that, under widely applicable regularity conditions, an estimator $\hat{\beta}_T$ for which

$$\sqrt{T}(\hat{\beta}_T - \beta) \mid \beta \xrightarrow[T\to\infty]{\mathcal{D}} N(0, \Sigma)$$

allows us with high accuracy and pre-sample probability, in large samples, to approximate the distribution of $\sqrt{T}(\beta - \hat{\beta}_T) \mid \hat{\beta}_T$ as $N(0, \Sigma)$. That is, we can interpret standard $(1 - \alpha)$ frequentist approximate confidence sets and regions generated from the frequentist asymptotic approximate distribution as if they were sets in parameter space with posterior probability $1 - \alpha$.

The regularity conditions that allow this simple conversion of frequentist confidence levels to Bayesian posterior probabilities do not hold universally, however. The most important condition is one that is usually needed also for generating asymptotically accurate frequentist confidence intervals: uniformity of the rate of convergence to the asymptotic distribution in some neighborhood of the true parameter value. In one important case, that of regression models for time series where there are some unit roots, the non-uniformity of convergence rates makes construction of frequentist asymptotic confidence regions difficult, while the usual normal model OLS computations of posterior probabilities continue to be asymptotically accurate under weak assumptions on disturbance distributions, even when unit roots are present (Kim, 1994).

These results allow us to form posterior distributions conditional on the estimator $\hat{\beta}_T$, as if we could see only the estimator and not the underlying data. This is of

course sometimes the case, as when we are reading an article that reports the estimators but not the full data set. A related result is that if we have a trustworthy model for which the likelihood function can be computed, the likelihood function will, under regularity conditions, take on a Gaussian shape in large samples, with the mean at the maximum likelihood estimator (MLE) and the covariance matrix given by the usual frequentist estimator for the covariance matrix of a MLE. [1] This result differs from the previous one in three important ways:

(i) It implies that conditioning on the MLE and using its asymptotic Gaussian distribution is, approximately in large samples, as good as conditioning on all the data.

(ii) It is only a computational shortcut: we can always check the accuracy of this Gaussian approximation to the shape of the likelihood by evaluating the likelihood itself at some points in the parameter space to see if the Gaussian approximation is accurate. In this respect, the result is very different from the computationally similar frequentist result, since there is no interpretation in the frequentist framework for the shape of the likelihood in any single finite sample.

(iii) This result has no "robustness" component. Frequentist asymptotic distributions often can be derived from weak assumptions that do not require specifying a full likelihood function. These weak assumptions carry over to the result that the frequentist asymptotic normal distributions can be "flipped" to be interpreted as conditional distributions of $\beta \mid \hat{\beta}_T$. But interpreting the likelihood shape as the posterior generally requires believing that the likelihood is correctly specified. We take up these robustness issues again in the section below on "sandwich" estimators.

So the view that Bayesian inference is harder is an artifact of the emphasis in the previous literature. Approximate frequentist distribution theory based on weak assumptions about distributions generally has an equally simple Bayesian interpretation that rests on equally weak assumptions. In fact, in one leading, practically important case (time series models with unit roots) the Bayesian approach leads to robust, convenient, asymptotically accurate distribution theory in a broader class of cases than does a frequentist approach.

Bayesians should recognize that the fundamental characteristic of their approach is that it conditions on the data, not that it insists on "doing things the hard way".

[1] See Gelman, Carlin, Stern, and Rubin (2004, Appendix B) for an informal discussion of these results, and the references given there for a more general treatment.

What I would call the "pragmatic Bayesian" approach accepts that in some circumstances we need to be able to reach conclusions conditional on standard, easily computed statistics, even when they are not sufficient statistics, and that sometimes approximate distribution theory, justified by the fact that in large samples with high probability the approximation is good, is the best that can be done with the available time and reources. Yet at the same time, it is always worthwhile to consider whether such compromises are throwing away a great deal of information or resulting in seriously distorted posterior distributions.

IV. Some less easily dismissed objections

IV.1. **Handy methods that seem un-Bayesian.** Probably more important than philosophical issues or computational costs to the persistence of frequentist inference in econometrics is the fact that some of the most widely used econometric methods seem to have no straightforward Bayesian counterpart. Instrumental variables (IV) estimation, "sandwich" style (or "clustered") robust estimated covariance matrices for least squares or IV estimates, generalized method of moments estimation (GMM), and kernel methods for estimating nonlinear regressions or pdf's all fall in this category. Each is easy to describe and implement. Each is justified by frequentist asymptotic theory that does not require that a correct likelihood be known; instead only a few moment conditions seem to be required. And none of them emerges neatly as the prescription for inference from a Bayesian approach.

We already have a quick answer to this gap in Bayesian methods — as we observed above, the approximate frequentist distribution theory for these estimators can be interpreted as allowing formation of approximate Bayesian posterior distributions, and this under essentially the same weak assumptions used in generating the frequentist asymptotic theory.

But every application of asymptotic approximate theory relying on "weak assumptions" involves a Bayesian judgment call. The asymptotics are invoked because accurate and convenient small-sample theory is not available under the weak assumptions. That is, though there may be models in the class defined by the weak assumptions under which the asymptotic theory is correct for the sample at hand, there are others for which the asymptotic theory is incorrect. If a frequentist uses the asymptotic approximation in a given sample, stating only the weak assumptions, he or she is implicitly ruling out those parts of the parameter space in which the asymptotic approximation, in this sample size, and with these conditioning variables, is inaccurate. These implicit assumptions are Bayesian in the sense that they invoke the researcher's pre-sample, or prior, beliefs about which parameter values or models are likely.

For example, it is widely understood that the distribution theory for ordinary least squares estimates that is exactly correct in the standard normal linear model (SNLM)

$$\underset{T\times 1}{y} \mid \underset{T\times k}{X} \sim N(X\beta, \sigma^2 I) \tag{1}$$

is approximately correct, asymptotically, under much weaker assumptions on the distribution of ε_t, albeit with some assumptions about the distribution of X added. For example, it is sufficient that the conditional covariance matrix of $y \mid X$ is independent of X and diagonal, the individual errors

$$\varepsilon_t = y_t - \underset{1\times k}{X_t}\beta \tag{2}$$

together with X_t form a jointly stationary, finite variance, i.i.d. process, and the minimal eigenvalue of $E[X_t' X_t]$ exceeds zero. But in a sample with given X matrix, we can always choose a zero-mean, finite-variance distribution for the ε_t's such that the Gaussian distribution theory is arbitrarily inaccurate. If we use the Gaussian theory in a sample of size 50, we are implicitly assuming that, say, a distribution for ε_t that puts probability .9999 on -1 and .0001 on 9999 is very unlikely. This is a distribution for ε_t that has zero mean and finite variance, but if it were the true error distribution and X included a constant vector, we would with high probability find a perfect fit in a sample of size 50, with the OLS estimates incorrect and the Gaussian distribution theory indicating no uncertainty at all about the parameter values. While it is easy to give examples like this of deviations from Gaussianity that in a particular sample would make the asymptotic theory inaccurate, it is not so easy to characterize the class of all such examples. Nonetheless applied work routinely claims not to require a normality assumption, without discussing how close to normality, and in what sense, the error distribution has to be in order to justify the asymptotic approximation.

Another example is kernel estimation of the nonlinear regression model, a simple form of which is

$$y_t = f(x_t) + \varepsilon_t \tag{3}$$

$$\varepsilon_t \mid x_t \sim N(0,1) \text{ for all } t \tag{4}$$

$$(y_t, x_t) \text{ independent of } (x_s, y_s)\,, \text{ all } t \neq s\,. \tag{5}$$

Kernel estimators estimate $f(x^*)$, the value of f at a particular point x, by averaging observed values of y_t for $x_t's$ within a neighborhood of length or volume h (called the bandwidth) around x^*. If h shrinks with sample size at an appropriate rate, then kernel estimates of f converge to true values and there is an asymptotic approximate distribution theory for the estimates. The asymptotic theory applies for any true f in

a broad class, subject to smoothness restrictions. In a finite sample, however, there will be a particular h in use, and it is clear that the estimates will be inaccurate if f varies greatly over regions of size h and also at points x^* for which there are few observed x_t values in an h neighborhood. In this case it is probably more intuitively obvious what implicit constraints are placed on the parameter space of f's when the asymptotic approximate distribution theory is invoked than in the case of OLS without explicit distributional assumptions. The fact that the restrictions are seldom brought out explicitly may therefore not matter so much, at least when bandwidth is reported prominently along with estimated f's.

But the facts that we can interpret these handy frequentist procedures to yield limited-information Bayesian posteriors and that when applied to small samples they in effect require asserting prior beliefs about parameter values do not form a completely satisfying Bayesian response. In assessing any statistical procedure and distribution theory, we would like to know whether it corresponds, at least approximately, to a full-information Bayesian posterior under some assumed model or class of models. Finding such models is useful in two ways. First, if the model does not exist, or turns out to require making clearly unreasonable assumptions, we may be led to ways of improving on the procedure. Second, if the model does exist it illuminates the assumptions implicit in the procedure and allows us to generate small-sample posterior distributions with weak-assumption asymptotic support. In the cases of the SNLM, GMM, IV, and OLS with sandwich-style distribution theory, there are models that are usable in small samples and have the frequentist asymptotic distributions as the limiting form of the posterior. In some cases these models are easy to use, in others not so easy. For kernel regression estimation, there apparently is no such model, though there is a class of models that delivers kernel-like estimates that adapt bandwidth to the density of the distribution of x values.[2]

These models are conservative in a sense — they promise no more precise estimates, asymptotically, than what can be obtained under the "weakest" of the models defined by the assumptions supporting the frequentist asymptotics. But to say they are conservative may be misleading. Using a model that fails to make efficient use of the data can lead to large, avoidable losses. Concluding that the data do not suport a finding that a drug or a social policy intervention has important effects is a costly error if the conclusion is incorrect, and reaching such a conclusion in a naive attempt to be "robust" by using "weak assumptions" is not in any real sense conservative.

IV.2. **Generating conservative models.** For a given set of maintained weak assumptions, there may be many finite-sample models that imply asymptotic likelihood

[2]See the discussion of nonparametric regression in Sims (2000).

shape that matches the shape of the frequentist asymptotic distribution of parameter estimates. But specifying even one of these models may at first glance seem a difficult task. There is a general principle that is often useful in this respect, however. Suppose we know the marginal distribution of the observed data y, given by a density function $q(\,)$, and the prior distribution on a parameter of interest β, given by $\pi(\beta)$. One appealing definition of a conservative model is then a conditional density for $y \mid \beta$, given by $f(y \mid \beta)$, that minimizes the Shannon mutual information between y and β subject to our weak assumptions. Mutual information is minimized at zero when y and β are independent, so if our weak assumptions do not rule out independence of y and β, the solution is $f(y \mid \beta) \equiv q(y)$, all y and β. But if the weak assumptions are enough to imply we can learn about β by observing y, we get a more interesting answer.

Suppose our weak assumption is that there is a vector $g(y, \beta)$ of functions such that $E[g(y, \beta) \mid \beta] = 0$, all β. The problem then is

$$\min_f \int \log\big(f(y \mid \beta)\big) f(y \mid \beta) \pi(\beta)\, dy\, d\beta - \int \log\big(q(y)\big) q(y)\, dy \tag{6}$$

$$\text{subject to } \int f(y \mid \beta) g(y, \beta)\, dy = 0\,, \text{ all } \beta\,, \tag{7}$$

$$\int f(y \mid \beta)\, dy = 1\,, \text{ all } \beta\,, \tag{8}$$

$$\int f(y \mid \beta) \pi(\beta)\, d\beta = q(y)\,, \text{ all } y\,, \tag{9}$$

$$f(y \mid \beta) \geq 0\,, \text{ all } y, \beta\,. \tag{10}$$

First order conditions then give us, when $f(y \mid \beta) > 0$ everywhere,

$$\pi(\beta)\Big(1 + \log\big(f(y \mid \beta)\big)\Big) = \lambda_1(\beta) g(y, \beta) + \lambda_2(\beta) + \lambda_3(y) \pi(\beta)\,, \tag{11}$$

and thus

$$f(y \mid \beta) = A(\beta) B(y) e^{\mu(\beta) g(y, \beta)}\,. \tag{12}$$

If we treat q as a choice variable rather than as exogenously given and add the corresponding constraint $q(y) = \int f(y \mid \beta) \pi(\beta)\, d\beta$, we conclude that $B(y)$ must be proportional to $q(y)$, y's marginal density.

Both with and without the constraint of a given $q(\,)$, this problem has the form of one with an objective function that is convex in f and constraints that are linear in f.[3] Therefore if we find a solution to the first-order conditions, it is a solution. Solving is not easy in general, however, because the $f > 0$ constraints may easily

[3]With q free, the objective is not convex in f and q jointly, but if q is substituted out as $q(y) = \int f(y \mid \beta) \pi(\beta)\, d\beta$, the objective is convex in f.

bind, so that no solution exists with $f > 0$ for all β, y. In fact it is easy to generate examples where the solution, when the distribution of y is unconstrained, makes the distribution of y discrete, in which case the FOC's for the continuously-distributed case have no solution and are little help. Nonetheless it is often possible to generate solutions for particular $\pi(), q()$ pairs, often for limiting cases where one or both are flat.[4] We shall see some examples of this below.

These solutions are in some respects natural and appealing. For the limiting case where $\pi(\beta)$ is flat, so that f is just the likelihood, they imply that the likelihood depends on y only through the function $g(y, \beta)$, and in the case where g is i.i.d., the likelihood for repeated samples $1, \ldots, T$ depends on y only through $\sum_t g(y_t, \beta)$. When $\partial g(y, \beta)/\partial y$ does not depend on β, the part of g that does not depend on β is a sufficient statistic. So if we can find a model like this its likelihood will lead to inference based on the same functions of the data that underly the frequentist distribution theory for the estimators.

However, models generated this way are conservative *only* for inference about β. If other aspects of the distribution of y are of interest, these models maximally conservative about β are generally dogmatic about other aspects of the distribution.

IV.3. **Non-parametrics.** In applied work, researchers often use distribution theory that claims to rely only on weak assumptions and think of themselves as thereby being robust against a wide array of deviations from some central model. Another approach to robustness is to contemplate specific deviations from the assumptions of a central model and experiment with extensions of the model that allow for these deviations. In regression models, for example, it is common to look for outliers, either informally or by expanding the model to allow for fat-tailed disturbance distributions or rare data points drawn from some very different distribution.

The conservative model approach of the previous section is a Bayesian analogue to the weak assumptions approach. Bayesians can also follow the other approach, more systematically than is usual in applied work: Put a prior distribution on an entire infinite-dimensional space of models. Converting model and prior to a posterior over the parameter space given the data is in principle a well-defined operation even in an infinite-dimensional space.

However, the mathematics of probability measures on infinite-dimensional spaces is in some respects paradoxical. In a finite-dimensional parameter space, if we use a prior distribution that has an everywhere positive density function, and if there

[4]The mathematical structure of this problem is dual to that of maximizing an objective function under uncertainty with a constraint on the mutual information between action and exogenous random disturbance. See Sims (2006).

is a consistent estimator for the unknown parameter, then the posterior distribution collapses on the true parameter value with probability one, except possibly on a subset of the parameter space of zero Lebesgue measure. It is a consequence of the martingale convergence theorem Doob (1949) that in any space Bayes estimates collapse on the true value with prior probability one. But Freedman (1963) and Diaconis and Freedman (1986) show that in infinite-dimensional parameter spaces, even if the prior puts positive probability on every open set, this is not enough to guarantee consistency of Bayesian posteriors on "most" of the parameter space.[5] Freedman characterizes the results as showing that in infinite-dimensional spaces Bayesian estimates are usually inconsistent. Many statisticians and econometricians think of these results as showing that Bayesian inference on infinite-dimensional parameter spaces is inherently unreliable. The theory of Bayesian inference on infinite-dimensional spaces, including both examples of inconsistency and discussion of priors that do guarantee consistency in important models, is laid out nicely in Gosh and Ramamoorthi (2003).

The Diaconis and Freedman results reflect some important general properties of infinite-dimensional linear spaces. Most infinite-dimensional parameter spaces are naturally taken to be complete, locally convex, topological vector spaces or large subsets of them. Examples are sequences of coefficients on lags in dynamic models, which are often naturally treated as lying in ℓ_2 (square summable sequences) or ℓ_1 (absolutely summable sequences) or spaces of functions in non-parametric regression models, where square-integrable functions (L_2) or absolutely integrable (L_1) might form natural parameter spaces. Such spaces are in important respects drastically different from finite-dimensional Euclidean spaces.

(1) Compact sets are nowhere dense, yet, as in any separable metrizable space, every probability measure puts probability one on a countable union of compact sets. Therefore every probability measure puts probability one on a set of Category 1, sometimes called a "meager" set.
(2) There is no analogue to Lebesgue measure — a measure that is translation invariant. In a finite-dimensional space, Lebesgue measure of a set S is the same as Lebesgue measure of $a + S$ for any vector a, and this property characterizes Lebesgue measure. No probability measure on an infinite dimensional space

[5]There is actually a series of papers by Freedman and Diaconis and Freedman working out various paradoxical, and not-so=paradoxical, examples.

is translation invariant. Worse than that, given any measure μ on an infinite-dimensional space, there is a set S with $\mu(S) > 0$ and $\mu(a + S) = 0$ for some a.[6]

(3) Metrics can agree in their definitions of convergence when restricted to finite-dimensional subspaces of the parameter space, yet disagree on the space as a whole.

The first two of these facts imply that any assertion of a probability distribution on an infinite-dimensional linear space is dogmatic, in the sense that it puts probability zero on "most" of the space, and controversial, in the sense that it puts probability zero on some sets that other distributions, differing only in where they are centered, give positive probability. The third fact implies that thinking about a prior on an infinite-dimensional space in terms of its implications for finite-dimensional subspaces can miss important aspects of its implications.

Bayesian inference on infinite-dimensional spaces is therefore always more strongly sensitive to the prior than is the case in finite-dimensional spaces. The practical implication of the complications of putting priors on infinite-dimensional spaces is that in non-parametric or semi-parametric Bayesian inference careful attention to the implications of the prior is much more important than in most finite-dimensional models. One could dismiss the inconsistency results by arguing that once one has carefully determined one's prior beliefs, failure of consistency on sets of prior probability zero should not be a concern. But the fact is that in practice priors are usually constructed to be convenient and made from conventional sets of building blocks (normal distributions, beta distributions, etc.). High-dimensional priors are often built up by considering low-dimensional subspaces first, then combining them. Priors cobbled together this way may sometimes roughly represent reasonable prior beliefs, but in high-dimensional spaces they can also easily turn out to be dogmatic in unintended ways. A failure of consistency on an intuitively large set is often a symptom of having specified a prior with unintended strong implications.

These pitfalls and paradoxes are not special defects of Bayesian inference. They reflect the difficulty of any kind of inference on infinite-dimensional spaces. The frequentist counterpart to the question of whether whether Bayesian posteriors are consistent is the question of whether there exists a sequence $\{S_n(X)\}$ of confidence sets such that for each sample size n, $P[\beta \in S_n(X) \mid \beta] \geq 1 - \alpha_n$, with $\alpha_n \to 0$ and the size of $S_n(X)$ shrinking to zero with n. An earlier paper of mine (1971) shows that obtaining asymptotically valid frequentist confidence intervals from sieve estimation schemes requires the same sort of strong a priori restrictions on the parameter space

[6]See Gelfand and Vilenkin (1964, Chapter 4, section 5.3) and Schaefer (1966, Chapter 1, section 4.3).

as are implied by assertion of a prior. For example, in the case of the non-parametric regression model (3-5), constructing asymptotically valid confidence intervals generally requires smoothness restrictions on the function $f(\cdot)$. But if we define the distance between f_1 and f_2 so that when f_1 and f_2 are close the distributions of the observed data they imply are close, we will find that $\int (f_1(x) - f_2(x))^2 dx$ is the right definition of closeness. With this definition of closeness (i.e. this topology), requiring even continuity of f eliminates most of the parameter space.

In time series applications we have models of the form $y_t = \sum_0^\infty \alpha_s \varepsilon_{t-s}$, where the ε_t are innovations in the y process. Sometimes we are interested in $\sum_0^\infty \alpha_s$, which determines the spectral density at zero frequency, or "long run variance". But the topology on α sequences induced by considering differences in implied distributions for y is ℓ_2, the metric of sum of squared deviations. $\sum \alpha_s$, though it is ℓ_1-continuous, is not ℓ_2-continuous. This means that we cannot get conclusions about $\sum \alpha_s$ from the likelihood alone: we have to eliminate most of the α space with prior restrictions. This is a well known issue[7]. For Bayesians, it implies that attempts to estimate long run variance are inevitably sensitive to the prior, even in large samples. For frequentists, it implies that the usual approach in time series models, estimating finitely parameterized models that expand with the amount of available data, makes conclusions about long run variance sensitive to the details of the finite-dimensional approximate models and the rate at which they are expanded. For estimation directly in the frequency domain, the usual kernel-smoothing approach yields estimates of long run variance that are sensitive to smoothness assumptions, as in kernel regression.

The mathematical theory underlying this section of the paper is difficult and abstract, but this should not be taken to mean that Bayesian inference on infinite (or even large) dimensional parameter spaces is impossible or even terribly difficult. It is no more difficult or pitfall-ridden than other approaches to the same models. But one has to be careful.

V. EXAMPLES

We consider several models and procedures that illustrate the principles and pitfalls described above.

V.1. **The Wasserman example.** Wasserman (2004) presents an example (11.9, p.186-188) meant to show that, because they are "slaves to likelihood", Bayesian methods "run into problems when the parameter space is high dimensional". It claims to display a case where any likelihood-based method must give poor results, while a simple, well-behaved frequentist estimator, together with confidence intervals, is

[7] Faust (1999) and Sims (1972) discuss versions of this issue.

available. The example does not actually support these claims. A simple estimator, uniformly better than the estimator suggested in the example, does emerge from a Bayesian approach based on a conservative model. And still better estimators are available if we recognize and model appropriately the presence of an infinite-dimensional unknown parameter.

We observe a sequence of i.i.d. draws (ξ_i, R_i, Y_i), $i = 1, \ldots, N$ of a vector of random variables (ξ, R, Y). These three random variables are jointly distributed with another, unobserved, random variable θ, and $\xi_i, R_i, Y_i, \theta_i$ are jointly i.i.d. across i. The joint distribution is specified as follows.

(1) ξ has a known distribution on $[0, 1]$.
(2) $R \mid (\ \xi, \theta)$ is 1 with probability ξ and zero with probability $1 - \xi$.
(3) $Y \mid (\ \xi, \theta, R = 1)$ is 1 with probability θ, 0 with probability $1 - \theta$.
(4) $Y \mid (\ \xi, \theta, R = 0)$ is not observed.

We do not know the joint distribution of (ξ, θ), only the marginal for ξ. We are interested in $\psi = E[\theta]$. The joint pdf of $\xi_i, R_i, Y_i, \theta_i$ is

$$p(\xi_i) q(\theta_i \mid \xi\ _i) \xi_i^{R_i} (1 - \xi_i)^{1 - R_i} \theta_i^{R_i Y_i} (1 - \theta_i)^{R_i - R_i Y_i} . \tag{13}$$

What we are trying to estimate is $\psi = \int \theta p(\xi) q(\theta \mid \xi\ )\, d\xi\, d\theta$, and the function $q(\cdot \mid \cdot\ )$ is an inifinite-dimensional unknown parameter.

Wasserman ignores the p and q components of the likelihood. That is, he writes down the conditional pdf of Y_i and $R_i Y_i$ given the random variables ξ_i and θ_i and calls that the likelihood. The fact that the distribution of ξ is assumed known means that $\xi(\omega)$, the function on the underlying space defining ξ_j as a random variable, is known. But this does not make ξ_i, the value of the random variable for the i'th observation, a parameter. The only unknown "parameter" here is $q(\theta \mid \xi\ )$, the conditional pdf for θ given ξ. Since θ_i is unobserved, to obtain a pdf for the observed data we need to integrate the pdf w.r.t. θ_i.

This is a challenging problem, because it involves estimation in the infinite-dimensional parameter space of unknown conditional distributions of $\theta \mid \xi$, yet it is of a type that arises in practice — some observations are missing, with the stochastic mechanism generating the missing cases uncertain.

What kind of prior probability measure over q would make sense here, for a Bayesian statistician reporting to an audience who are interested in ψ but uncertain of its value? It perhaps goes without saying, that we can immediately rule out priors that imply $E[\theta]$ is known with certainty, or even with great precision. If we are interested in estimating ψ and uncertain about it, we can't have a prior that implies we know it a priori. (This point may seem pedantic, but as we will see below, it is relevant to Wasserman's discussion of the problem.)

Independence. It would be convenient analytically to postulate that θ and ξ are independent, i.e. that $q(\theta \mid \xi\ )$ does not depend on ξ. If that were true, we could focus attention on the $R_i = 1$ cases and treat them as an ordinary i.i.d. sample with $P[Y_i = 1] = \theta_i$. The selection mechanism (the ξ_i's) would be irrelevant to inference. A simple Bayesian setup for this case would postulate that $\theta \sim \text{Beta}(\psi\nu, (1-\psi)\nu)$, where $\nu > 0$ is a constant that determines how tightly we expect observed θ values to cluster around their mean, ψ. We would have to provide a prior for ψ also, which might well be taken as uniform on $[0,1]$. This would lead to a $\text{Beta}(n+1, m+1)$ posterior distribution on ψ, where n is the number of observed $Y_i = 1$ cases and m is the number of observed $Y_i = 0$ cases. The posterior mean of ψ would therefore be $\tilde{\psi} = (n+1)/(n+m+2)$. For very small numbers of observed Y's, this is close to the prior mean of .5, but for large n and m it is close to $n/(n+m)$, an intuitively simple and appealing estimator for this independence case.

The posterior distribution is not only nicely behaved here, it is in a certain sense robust. We know that in this i.i.d. case the sample mean of the Y_i's, which is negligibly different from $\tilde{\psi}$ in large samples, has a Gaussian asymptotic distribution with variance consistently estimated by $1/N$ times the sample variance. If we were to take a "Bayesian limited information" approach (Kwan, 1998) and construct an approximate posterior based just on this frequentist asymptotic theory, we would get exactly the same limiting distribution that emerges from using the exact Bayesian small-sample theory underlying $\tilde{\psi}$. In other words, we know that any other model that leads to focusing attention on sample mean and variance of the observed points will lead to about the same inferences in large samples.

Dependence: direct approach. However, all we have done here is to observe that under independence we can easily "integrate out" the θ_i's, so the problem becomes finite-dimensional. The reason for introducing ξ in the first place has to be that we are concerned that it might not be independent of θ. That is, we must think it possible that observations with particular values of θ may be more or less likely to be selected into the sample. If, for example, high θ's tend to correspond to high ξ's, we will be observing Y's for cases with unrepresentatively high values of θ and $\tilde{\psi}$ will be biased upward.

If we want to avoid such bias, and at the same time generate reliable statements about the uncertainty of our estimates, there is no choice (whether one takes a frequentist or a Bayesian approach) but to model the joint behavior of θ and ξ. The unknown parameter here is $q(\theta \mid \xi\ )$, the conditional pdf of θ. As in the independence version of the problem, mistakes on the form of q as a function of θ for given ξ probably don't matter much, so long as we have parameters for location and scale.

The crucial thing is to model flexibly the dependence of q on ξ. Here, as often in infinite-dimensional problems, a natural and convenient approach is a "sieve". That is, we think up a countable sequence of finite-parameter models, indexed by k, in such a way that all the shapes for q as a function of θ that we think likely are well approximated by some element of the k'th parameter space, if k is large enough. A Bayesian sieve puts probability over the k's as well as over the parameters within each space. Such a setup leads generally to, in a given sample, putting nearly all probability on a small number of k values (often just one), yet provides a systematic rule for increasing the size of the parameter space as evidence of more complex behavior accumulates. There are many ways to set up reasonable sieves in this case. A straightforward approach might be to partition $[0,1]$ into k intervals I_ℓ, $\ell = 1,\ldots,k$, and postulate that there is a fixed q_ℓ that holds for all $\xi \in I_\ell$. If we use the same exchangeable-Beta form within each I_ℓ that we used above for the independence case, but of course with mean ψ_ℓ varying with ℓ, we get k replicas of the independence case problem. This will, at the expense of a bit more work, insulate us against bias from dependence between ξ and θ.

Of course this sieve approach puts prior probability one on step-function forms for q as a function of ξ. Even though these forms approximate a very wide class of functions well, they are not in themselves a large fraction of the space of all possible forms for q as a function of ξ. As we have pointed out above, this is a problem for any (Bayesian or frequentist) approach to inference on an infinite-dimensional topological vector space. There is no way to make reliable probability statements about estimates in such a space without concentrating attention on a subset of the space that is small, in the same sense that a countable union of finite-dimensional spaces is a small subset of, say, the space of square-summable sequences. We can hope that estimators are consistent on the whole space, and the subset on which statements about uncertainty concerning a finite-dimensional parameter of interest behave well may be much larger than the support of the prior distribution. But generally we can't proceed in such a way as to rule out the possibility of repeatedly claiming arbitrarily high certainty about assertions that aren't true, if we insist that every parameter value in an infinite-dimensional topological vector space is a possibility.

Limited information. The Horwitz-Thompson estimator that Wasserman suggests is a simple, information-wasting approach. It ignores the difference between observations in which Y is unobserved and observations in which Y is observed to be zero, proceeding as if all we observed were 0's when either $Y_i = 0$ or $R_i = 0$ and $Z_i = Y_i/\xi_i$ when $R_i = Y_i = 1$. The Horwitz-Thompson estimator simply estimates

ψ as $\hat{\hat{\psi}} = \sum Z_i/N$, where N is the full sample, including all the zero values for Z_i. Supposing the Z_i's were all we could see, what would be a Bayesian approach?

We would have i.i.d. draws of Z_i, which has discrete probability at 0 and an unknown distribution over $(1, \infty)$ conditional on being non-zero. The probability of a non-zero observation is $\alpha > 0$. If we maximize entropy of Z subject to the known mean, we arrive at an exponential distribution for $Z - 1$ with rate parameter μ. The likelihood is then

$$\alpha^n (1-\alpha)^m \mu^n e^{-\mu \sum_{Z_i>0}(Z_i-1)}, \tag{14}$$

where n is the number of non-zero observations and m the number of zero observations on Z. Note that $1 > E[Z] = \alpha(\mu+1)/\mu$, which in turn implies $\mu > \alpha/(1-\alpha)$. One class of conjugate prior is, for $\gamma > 0$, proportional to $\gamma e^{-\mu\gamma}$ over the region where $0 < \alpha < \mu/(\mu+1)$. Note that because of the bound on the range of α for given μ, this prior, though "flat" in α conditional on μ, has a marginal pdf proportional to $\exp(-\gamma\alpha/(1-\alpha))$, which declines sharply as $\alpha \to 1$.

If we ignore the inequality constraint on (α, μ), we can compute analytically the posterior mean of $\theta = E[Z_i] = \alpha \cdot (\mu^{-1} + 1)$. With the exponential prior on μ, the posterior makes α and μ independent. The α posterior is Beta(n, m), while μ is Gamma$(n+1, (n+m)(\bar{Z} - n/(n+m) + \gamma/(n+m)))$, where $\bar{Z}$ is the sample mean of the Z_i's, averaging over all Z's, zero and non-zero. Since the expectation of the inverse of a Gamma(p, a) variable is $a/(p-1)$, the posterior mean of θ is then

$$\hat{\psi} = \frac{n}{n+m}\left(\frac{(n+m)(\bar{Z} - n/(n+m) + \gamma/(n+m))}{n} + 1\right) = \bar{Z} + \frac{\gamma}{n+m}.$$

This is very close, for large n and m, to simply using the sample mean of Z as the estimate of θ.

However, it is not a good idea to ignore the inequality constraint. If one does so, the estimator can emerge as greater than one, indeed is quite likely to do so if θ's distribution concentrates near its upper limit of 1. It doesn't seem that there is an easy analytic formula for the posterior mean that accounts for the inequality constraint, but the problem reduces to a simple one-dimensional numerical integration.

As a simple example consider the case where the mean of a sample of 100 draws of Z_i is .8. In Figure 1 we show the posterior pdf's and $\hat{\psi}$ for the cases where the number of non-zero draws is 3, 10, 50 and 70. These pdf's were estimated from 4000 Monte Carlo draws from each posterior distribution. Draws were made from the posterior ignoring the inequality constraint (which makes α and μ independent Beta and Gamma variates, respectively) and those violating it were discarded. It is clear that, as makes intuitive sense, a sample with 70 nonzero Z_i draws averaging 8/7 (and hence clustered close to the $Z_i > 1$ lower bound for nonzero draws) gives

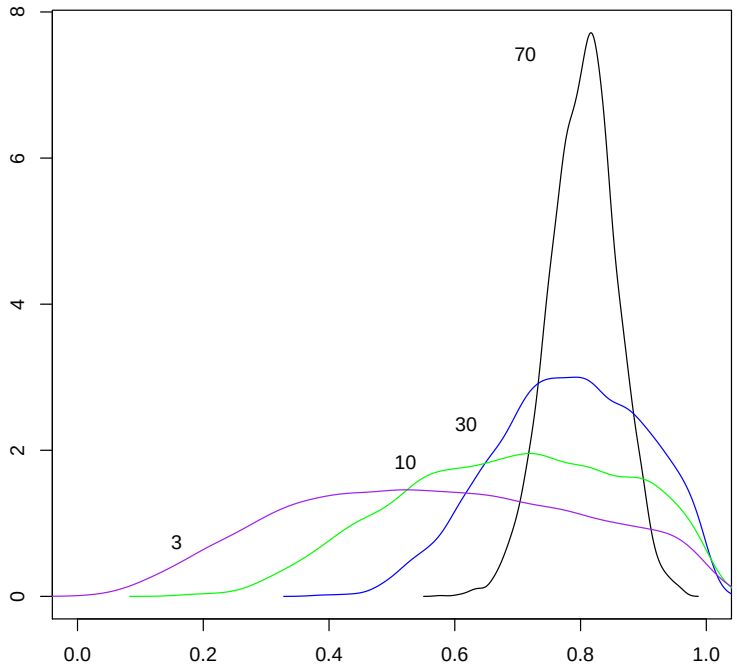

FIGURE 1. Posterior densities for ψ with $\hat{\hat{\psi}} = .8$, various numbers of non-zero Z

much more reliable inference about ψ than a sample with the same sample mean for Z, but with only three nonzero values, averaging about 27.

Of course the Horwitz-Thompson estimator is $\hat{\hat{\psi}} = .8$ for all of these samples. Wasserman points out that there is a uniform upper bound on the variance of $\hat{\hat{\psi}}$, if we know an a priori certain positive lower bound on ξ, say $\xi > \delta > 0$. The bound is $1/(N\delta) - 1/N$ if $\delta < .5$ and $1/(4N\delta^2)$ if $\delta > .5$[8] Because these are uniform across the parameter space, they can be used to construct (very) conservative interval estimates. In a sample this size, any value of δ less than about $\frac{1}{6}$ produces conservative interval estimates of this type that include the entire $(0, 1)$ interval.

Wasserman's conclusions about Bayesian methods in this example. Wasserman argues that Bayesian posterior distributions for this problem must be nearly identical to

[8]Wasserman suggests the weaker lower bound $1/(N\delta^2)$, which avoids using separate formulas for the two ranges.

priors. He also argues that Bayesian estimators for ψ must ignore observations on ξ, because the likelihood factors into a function of ξ_i's and a function of θ_i's. He also says there is no way to derive the Horwitz-Thompson estimator from a Bayesian perspective.

The claim that Bayesian approaches must make the posterior insensitive to the data arises from Wasserman's having implicitly assumed a prior drawn from a class that makes ψ known a priori with near certainty — recall that we began by noting that this was obviously a bad idea. However it is not quite so obviously a bad idea in the specific setup that Wasserman uses. Our presentation of the example makes ξ and θ continuously distributed. Probably to avoid in a textbook having to talk about "parameters" that are functions on the real line, Wasserman supposes that ξ and θ are random variables defined on a large, but finite and discrete, probability space, whose points are indexed by the integers $\omega = 1, \ldots, B$, with all points ω having measure $1/B$. He suggests thinking of B as much larger than sample size N. As a practical matter, if, say, $B = 100,000$ and $N = 1000$, there is little difference between a sample $\{\theta_i\}$ drawn as N i.i.d. draws from $f(\psi, \nu)$ and a sample drawn in two stages — first by drawing B values of θ from $f(\psi, \nu)$, then drawing N values by sampling with replacement from the initially drawn B values of θ. If ψ is the expectation of a random variable with the f pdf, the posterior on ψ after N observations will differ slightly from the posterior on $(1/B)\sum_1^B \theta(\omega)$. The Bayesian methods discussed above can be extended straightforwardly to produce estimators of both these concepts.

But when Wasserman writes the likelihood in terms of $\theta(\omega)$, $\omega = 1, \ldots, B$ as "parameters", he notes that most of these parameters don't appear in the likelihood function and that the likelihood factors into one piece involving these parameters and another piece involving the $\xi(\omega)$'s. He then concludes that Bayesian methods must have posterior means equal to prior means for most of the $\theta(\omega)$ values, since most of them will correspond to ω's that have not been observed. But, as should be clear by now, this is just a mistake. Priors don't have to preserve independence of parameters that would be independent if we used the likelihood directly as posterior (i.e. used a "flat" prior). It makes no sense in this problem to assume prior beliefs about θ independent across ω values, as this would imply that we knew ψ with extremely high precision a priori. It also makes no sense to assume independence in the prior of ξ and θ, as that would imply that we knew selection was independent of the parameter of interest, and hence that we could ignore it in conducting inference.

So one might say that this section's analysis has turned Wasserman's conclusions completely on their ear. But this is not quite true. Wasserman is an excellent statistician who understands and teaches Bayesian methods. The structure of this tricky

little problem led him to postulate a prior that, without his apparently realizing it, was unreasonably dogmatic along two dimensions important to the data analysis (independence of $\theta(\omega)$ across ω and independence between ξ and θ). Asymptotically, with B held fixed and sample size increasing to infinity, priors do not affect inference, in this problem as in many others. Indeed if B were not extremely large allowing for dependence across i in the prior for θ's and dependence between ξ and θ might have little effect on inference. It is the large B that makes these apparently ordinary assumptions influential, and unreasonable. So what Wasserman's example illustrates is one of the pitfalls of inference in high-dimensional spaces — priors are inevitably dogmatic in certain directions, and it is important to insure that they are not dogmatic in ways that matter to the inferential problem at hand.

The example also illustrates the fact that apparently reasonable methods derived by non-Bayesian methods sometimes turn out to be unexpectedly quirky, because there are types of samples or points in the parameter space where they show surprising misbehavior. The unbiased, consistent, Horwitz-Thompson estimator turns out to be inadmissible, meaning strictly dominated in mean squared error, because it will occasionally produce $\hat{\hat{\psi}} > 1$, and even do so with probability near .5 at some points in the parameter space. Of course reasonable applied researchers would not accept $\hat{\hat{\psi}}$ estimator values outside $[0, 1]$, and would adjust them. But if they were the usual sort of practical frequentist, they might well not explain, in those instances where adjustment is *not* needed, that because they *would* modify $\hat{\hat{\psi}}$ if the estimates exceeded 1, the estimates they are presenting are not in fact unbiased.

V.2. **Robust variance estimates in regression.** Ordinary least squares (OLS) estimates of a linear regression equation are in a certain sense robust. It is easily verified that the standard normal linear model (SNLM) emerges as the solution to the minimum mutual information problem when the moments assumed are $E[y \mid X] = X\beta$ and $\text{Var}(y \mid X) = \sigma^2$, and this model of course yields the same distribution theory that is obtainable as an asymptotic distribution under much weaker assumptions. Nonetheless it has become common in applied work, since computer packages have made this a one-click option, to present OLS estimates with "clustered" or "robust" standard errors rather than those that emerge from the SNLM. Müller (2009) argues that a Bayesian decision-maker can justify using OLS with a sandwich covariance matrix when the probability limit of the OLS estimator is the object of interest, despite the fact that the SNLM is known not to be the true model.

Müller's suggestion can be supported by a limited-information Bayesian argument. If the data are i.i.d., the frequentist asymptotic distribution theory that asserts

that in large samples approximately

$$\hat{\beta}_{OLS} \sim N(\beta, (X'X)^{-1} \sum X_t' X_t \hat{u}_t^2 (X'X)^{-1}) \tag{15}$$

can be flipped to assert that the posterior, conditional on $\hat{\beta}$ and the sandwich covariance estimate itself, has this same form.

But if the sandwich covariance matrix differs much from the usual $\hat{\sigma}^2(X'X)^{-1}$ covariance matrix for β, this must reflect misspecification in the SNLM model underlying the OLS estimate. Such a discrepancy can be regarded as a test of misspecification, therefore, and Müller (and others) have suggested interpreting it this way. It might be, then, that whenever the sandwich covariance matrix gives different results, it should itself be regarded as only a temporary patch. The best reaction to it, subject to time and computaional constraints, is to look for a better model in which the sandwich covariance matrix and the likelihood-based one would coincide.

This thought leads to an interesting question: Are there classes of models in which, in large samples, there is little to be gained from trying to use the correct model's likelihood rather than OLS plus the robust covariance matrix? This question is relevant both because we might sometimes actually be using such models and also because it might turn out that such models do not make sense in a given application, so that extra effort to improve on OLS plus sandwich is expected have a large payoff.

The sandwich estimator for this i.i.d. OLS case is often characterized as "hetroskedasticity-robust". It is clear that if $E[u_t^2 \mid X_t] = \sigma^2$, so there is no heteroskedasticity that depends on X_t, the sandwich and the usual covariance estimator asymptotically coincide. It may therefore seem that the situation where one might as well be content with the sandwich estimator is that in which there is conditonal heteroskedasticity, $E[u_t^2 \mid X_t] = \sigma^2(X_t)$, while the assumption from the SNLM that $E[u_t \mid X_t] = X_t\beta$ is maintained. But this is incorrect.

It is an implication of the results in Chamberlain (1987) that the sandwich estimator provides the efficiency bound for the OLS estimator when estimation is based only on the assumption that $E[X'y] = E[X'X]\beta$. That is, this bound applies when the parameter of interest β is the population value of the least squares fit. This does not mean that when this moment condition is satisfied no estimator can have a lower asymptotic variance. If, say, $E[y \mid X] = X\beta$ and $u = y - X\beta \mid X$ has the double-exponential pdf $.5\alpha e^{-\alpha}$, a minimum absolute deviations estimator will improve on OLS and have lower asymptotic variance. The efficiency bound means that no estimator can improve on the bound without restricting the space of possible models a priori. Or, to put it another way, any estimator that improves on the bound for some particular model must do worse than the bound on some other model that satisfies the moment restrictions.

If we restrict the space of models so that they satisfy not only $E[X_t' y_t] = E[X_t' X_t]\beta$, but also the stronger requirement that $E[y_t \mid X_t] = X_t\beta$, the sandwich covariance is no longer the efficiency bound. In this case the efficiency bound is achieved by the weighted least squares estimate, with the usual normal model estimated covariance matrix. And in the special case where $\text{Var}(y_t \mid X_t) \equiv \sigma^2$, the usual SNLM distribution theory corresponds to the efficiency bound. In the homoskedastic case, as Müller points out, the efficiency bound implies that any attempt to model the distribution of disturbances can improve inference only in small samples, or else for an a priori limited range of true distributions of the disturbances: In a large enough sample, any general model for the disturbance distribution must converge to OLS with the efficiency bound as covariance matrix.

But this discouraging result of limited payoff to modeling disturbances does not correspond to any limited payoff from modeling conditional heteroskeasticity. The efficiency bound does *not* imply that any attempt to model the form of $\sigma^2(X_t)$ must in large samples either produce worse estimates for some possible models satisfying the moment restrictions or else produce the sandwich limiting distribution. When heteroskedasticity is present, there is likely to be a gain in modeling it and using weighted least squares rather than OLS. This leaves open the possibility (which might be worth investigating further) that there are classes of priors for the $\sigma^2(X_t)$ function that would produce gains in efficiency over most of a general space of $\sigma^2()$ functions.

So do we conclude that OLS with the sandwich is always a lazy shortcut, never an approach that an energetic Bayesian should be content with, even in large samples? No. The assumption $E[y_t \mid X_t] = X_t\beta$ is restrictive, and there are conditions under which we might not want to impose it. This is the situation in which there is a nonlinear regression function $E[y_t \mid X_t] = f(X_t)$, so that the residual $u_t = y_t - X_t\beta$ has non-zero mean conditional on X. Here β is defined as the set of coefficients that provide the best *linear* fit in predicting y_t from X_t. Since the true regression function is nonlinear, we cannot determine the best linear fit from knowledge of the nonlinear regression function $f()$ alone; it depends also on the distribution of X_t. It is this dependence on the unknown distribution of X_t that forces us to the sandwich covariance matrix as the asymptotic efficiency bound.

Chamberlain's original paper proved its results by starting with the case of discrete joint distributions of y_t, X_t — that is, the case where there are only a finite number of possible y_t, X_t pairs, from which all sample values are drawn. Recently Szpiro, Rice, and Lumley (2008) have provided a Bayesian dual to Chamberlain's result. They have shown that in a model with discretely distributed X_t and possibly nonlinear $E[y_t \mid X_t]$, the Bayesian posterior distribution for β asymptotically takes

on the form of a normal distribution centered at OLS with the sandwich covariance matrix. They conjecture that the result would carry over to smoothly distributed X_t, so long as the distribution is unknown and has to be estimated.

So what should we conclude about the pragmatic Bayesian attitude toward OLS with clustered or "heteroskedasticity-consistent" standard errors? If presented with such estimates, but not the full sample, they can be given an approximate Bayesian interpretation by flipping the asymptotic distribution. If presented with the full sample, in a context where an estimate of $f(X_t) = E[y_t \mid X_t]$ is needed and the sandwich covariance matrix is quite different from the SNLM's covariance matrix around the OLS estimate, deviations from the SNLM in two directions should be considered. One possibility is that $f()$ is nonlinear, in which case nonparametric regression or expanding the parameter space to allow nonlinear terms in X_t into the regression is appropriate. In this case obviously OLS in large samples is unjustifiable. The other possibility is that there is reason to rely on the assumption that f is linear. In that case modeling and estimating the $\sigma^2(X_t)$ function and replacing OLS by weighted least squares makes sense.

Only in the case where a nonlinear f is likely, the sample is large, and an estimate of the best linear predictor for y_t based on X_t is needed, should a pragmatic Bayesian conclude that, in a large sample, there is likely little return to replacing OLS+sandwich with an explicit likelihood-based analysis. Cases where such a linear estimate is useful despite (because of the large sample) it being clear that the estimate is biased over some ranges of X values are probably rare.

VI. Conclusion

References

Chamberlain, G. (1987): "Asymptotic Efficiency in Estimation with Conditional Moment Restrictions," *Journal of Econometrics*, 34, 305–334.

Diaconis, P., and D. Freedman (1986): "On the Consistency of Bayes Estimates," *The Annals of Statistics*, 14(1), 1–26.

Doob, J. (1949): "Application of the theory of martingales," in *Colloque Internationale du Centre Nationale de Recherche Scientifique*, pp. 22–28, Paris.

Faust, J. (1999): "Conventional Confidence Intervals for Points on Spectrum Have Confidence Level Zero," *Econometrica*, 67(3), 629–37.

Freedman, D. A. (1963): "On the Asymptotic Behavior of Bayes' Estimates in the Discrete Case," *The Annals of Mathematical Statistics*, 34(4), 1386–1403.

Gelfand, I., and N. Y. Vilenkin (1964): *Generalized Functions, Volume 4, Applications of Harmonic Analysis*. Academic Press, New York, London.

GELMAN, A., J. B. CARLIN, H. S. STERN, AND D. B. RUBIN (2004): *Bayesian Data Analysis*. Chapman and Hall, London, 2nd edn.

GOSH, J., AND R. RAMAMOORTHI (2003): *Bayesian Nonparametrics*. Springer-Verlag, New York.

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

KIM, J. Y. (1994): "Bayesian Asymptotic Theory in a Times Series Model with a Possible Nonstationary Process," *Econometric Theory*, 10(3), 764–773.

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

MÜLLER, U. K. (2009): "Risk of Bayesian Inference in Misspecified Models and the Sandwich Covariance Matrix Estimator," Discussion paper, Princeton University.

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.

SCHAEFER, H. H. (1966): *Topological Vector Spaces*. Macmillan, New York.

SIMS, C. A. (1971): "Distributed Lag Estimation When the Parameter-Space is Explicitly Infinite-Dimensional," *Annals of Mathematical Statistics*, 42(5), 1622–1636.

——— (1972): "The Role of Approximate Prior Restrictions in Distributed Lag Estimation," *Journal of the American Statistical Association*, 67(337), 169–175.

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

——— (2006): "Rational Inattention: Beyond the Linear-Quadratic Case," *American Economic Review*, 96(2), 158–163, Proceedings issue.

SZPIRO, A. A., K. M. RICE, AND T. LUMLEY (2008): "Model-Robust Regression and a Bayesian 'Sandwich' Estimator," UW Biostatistics Working Paper Series 338, University of Washington University, http://www.bepress.com/uwbiostat/paper338.

WASSERMAN, L. (2004): *All of Statistics: A Concise Course in Statistical Inference*, Springer Texts in Statistics. Springer.

DEPARTMENT OF ECONOMICS, PRINCETON UNIVERSITY
E-mail address: `sims@princeton.edu`