USING A LIKELIHOOD PERSPECTIVE TO SHARPEN ECONOMETRIC DISCOURSE: THREE EXAMPLES

CHRISTOPHER A. SIMS

Abstract. This paper discusses a number of areas of inference where dissatisfaction by applied workers with the prescriptions of econometric high theory is strong and where a likelihood approach diverges strongly from the mainstream approach in its practical prescriptions. Two of the applied areas are related and have in common that they involve nonstationarity: macroeconomic time series modeling, and analysis of panel data in the presence of potential nonstationarity. The third area is nonparametric kernel regression methods. The conclusion is that in these areas a likelihood perspective leads to more useful, honest and objective reporting of results and characterization of uncertainty. It also leads to insights not as easily available from the usual perspective on inference.

1. Introduction

Many econometricians are committed, at least in principle, to the practice of restricting probability statements that emerge from inference to pre-sample probability statements—e.g. "If I did this analysis repeatedly with randomly drawn samples generated with a true β of 0, the chance that I would get a $\hat{\beta}$ as big as this is .046." Of course if we are to make use of the results of analysis of some particular data set, what we need is to be able to make a post-sample probability statement—e.g., "Based on analysis of these data, the probability that β is 0 or smaller is .046." The latter kind of statement does not emerge as an "objective" conclusion from analysis of data, however. If this is the kind of probability statement we aim at, the objective part of data analysis is the rule by which probability beliefs we held before seeing the data are transformed into new beliefs after seeing the data. But providing objective analysis of data that aims to aid people in revising their beliefs is quite possible, and is the legitimate aim of data analysis in scholarly and scientific reporting.

Date: November 24, 1998.

Copyright 1998 by Christopher A. Sims. This version may be reproduced for educational and research purposes, so long as this copyright notice is included and the copies are not sold, even to recover costs.

Pre-sample and post-sample probabilities are often closely related to each other, requiring the same, or nearly the same calculations. This is especially likely to be true when we have a large i.i.d. sample and a well-behaved model. This is one reason why the distinction between pre- and post-sample probability hardly ever enters the discussion of results in natural science papers. But in economics, we usually have so many models and parameters potentially available to explain a given data set that expecting "large" sample distribution theory to apply is unrealistic, unless we artificially restrict formal analysis to a small set of models with short lists of parameters. Econometric analysis generally does make such deliberate artificial restrictions. Applied econometricians and users of their analyses understand this as a practical matter and discount the formal probability statements emerging from econometric analyses accordingly. If the discounting is done well, the result need not be badly mistaken decisions, but if the formal probability statements themselves can make econometricians look foolish or hypocritical. It would be better if econometricians were trained from the start to think formally about the boundaries between objective and subjective components of inference.

In this paper, we are not going to expand further on these broad philosophical points. Instead we are going to consider a series of examples in which a Bayesian approach is needed for clear thinking about inference and provides insights not easily available from standard approaches.

2. Possibly Non-Stationary Time Series

2.1. **Why Bayesian and Mainstream Approaches Differ Here.** Sample information about variation at frequencies with wavelength in the neighborhood of T in a sample of size T is inherently weak—only about one instance of a cycle associated with such a wavelength can have been observed. We cannot expect that sample information will dominate prior beliefs about such variation, whether those beliefs are formulated explicitly or enter the analysis through the back door in the form of conventional modeling assumptions. Bayesian approaches that make the role of prior information explicit are therefore in conflict with mainstream approaches that attempt to provide apparently objective rules for arriving at conclusions about low-frequency behavior.

Because mainstream approaches do not use probability to keep track of the role of non-sample information in determining conclusions about low frequency behavior, they can lead to unreasonable procedures or to unreasonable claims of precision in estimates or forecasts. When forecasting or analysis of behavior at low frequencies is the center of interest in a study, good research

practice should insist on modeling so that the full range of a priori plausible low-frequency behavior is allowed for. Of course this is likely to lead to the conclusion that the data do not resolve all the important uncertainty, so that the model's implied probability bands will (accurately) appear wide and will be sensitive to variations in auxiliary assumptions or (if explicitly Bayesian methods are used) to variations in the prior.

But mainstream approaches instead tend to lead to an attempt to find a model that is both simple and "acceptable"—not rejected by the data—and then to the making of probability statements conditional on the truth of that model. Parameters are estimated, and ones that appear "insignificant", at least if there are many of them, are set to zero. Models are evaluated, and the one that fits best, or best balances fit against parsimony, is chosen. When applied to the problem of modeling low frequency behavior, this way of proceeding is likely to lead to choice of some single representation of trend behavior, with only a few, relatively sharply estimated, free parameters. Yet because the data are weakly informative, there are likely to be other models of trend that fit nearly as well but imply very different conclusions about out-of-sample low frequency behavior.

The Bayesian remedy for this problem is easy to describe. It requires, for decision-making, exploring a range of possible models of low frequency behavior and weighting together the results from them on the basis of their ability to fit the data and their a priori plausibility. The exact weights of course depend on a prior distribution over the models, and in scientific reporting the aim will be to report conclusions in such a way that decision-makers with differing priors find the report useful. Where this is feasible, the ideal procedure is simply to report the likelihood across models as well as model parameters.

Good mainstream econometric practice can, by reporting all models tried that fit reasonably well, together with measures of their fit, and by retaining "insignificant" parameters when they are important to substantive conclusions, give an accurate picture of the shape of the likelihood function.

Often, though, "reporting the likelihood" is not feasible, because there are too many parameters or (equivalently, since choice among models is estimation of a discrete parameter) too many models. Here the Bayesian approach will involve integrating out parameters that are not of central interest. Doing this requires use of a prior, however. Unlike reporting the likelihood, which produces information usable by people with varying priors, reporting a marginalized likelihood or marginalized posterior may not be helpful to people whose priors do not match the prior used in performing the integration (even if the

likelihood itself has been integrated, implying a flat prior). Good reporting, then, will require doing a good job of choosing one or more priors that come close to the beliefs of many readers of the report. Reporting marginal posteriors under a range of priors can be thought of as describing the likelihood by reporting values of linear functionals of it rather than its values at points.

To this point we have simply applied to the context of modeling nonstationary data general principles concerning the difference between pre-sample and post-sample forms of inference. But there are some specific ways in which "classical" approaches to inference either lead us astray or hide important problems from view.

2.2. **Bias Toward Stationarity.** In a classic paper in econometric theory, Hurwicz (1950) showed that least squares estimates of ρ in the model

$$\begin{aligned} y(t) &= \rho y(t-1) + \varepsilon(t) \\ \varepsilon(t) &| \{y(s), s < t\} \sim N(0, \sigma^2) \end{aligned} \tag{1}$$

is biased toward zero. I think most people interpret this result to mean that, having seen in a given sample the least squares estimate $\hat{\rho}$, we should tend to believe, because of the bias toward zero in this estimator, that probably the true value of ρ is greater than $\hat{\rho}$, i.e. that in thinking about the implications of the data for ρ we need to "correct" the biased estimator for its bias. As Sims and Uhlig (1991) showed, this is not true, unless one started, before seeing the data, with strong prior beliefs that ρ values near 1 are more likely than smaller ρ values. The reason is that the bias is offset by another effect: $\hat{\rho}$ has smaller variance for larger ρ's, a fact that by itself would lead us to conclude that, having seen the sample data, the true ρ is probably below $\hat{\rho}$.

In this context, then, the presample probability approach gives us a misleading conclusion. But though bias in this model does not have the implications one might expect, there is a serious problem with naive use of least squares estimates in more complex autoregressive time series models, and classical asymptotic theory gives us misleading answers about it.

2.3. **Implausible Forecasting Power in Initial Conditions.** In a model with a constant term as well, or worse, with a constant and a linear trend term, the chance of a sample from a nonstationary or near-nonstationary process imitating a stationary process is higher, and also the chance of explaining a lot of variation as emerging deterministically from initial conditions rises. When this happens, the results usually do not make sense. This situation is reflected in the drastically increased (presample) bias when constant terms or

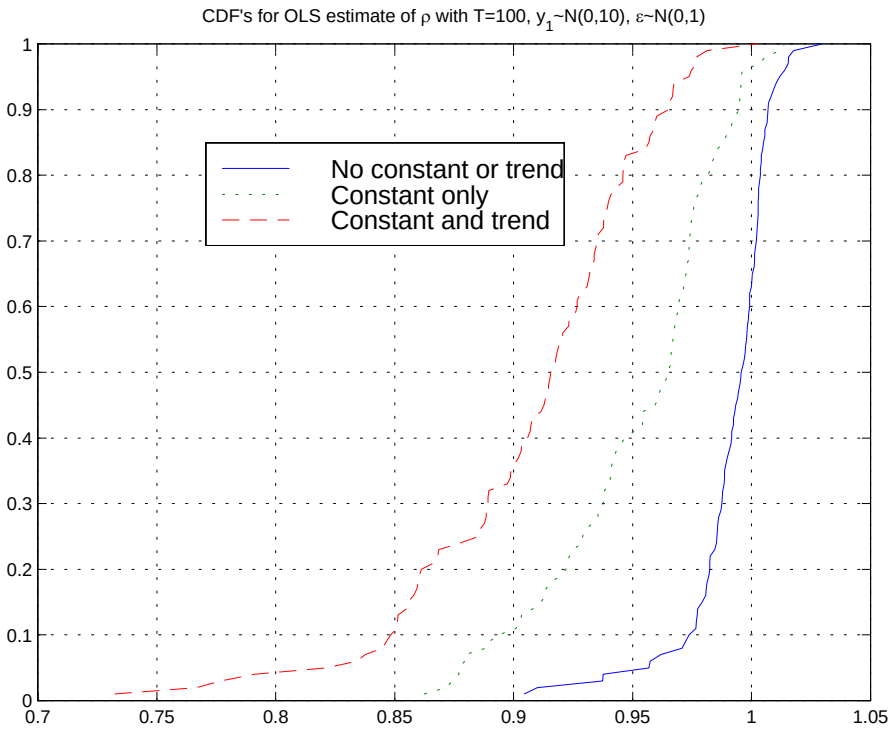

FIGURE 1. CDF's of OLS $\hat{\rho}$ when $\rho = 1$

constants with trends are added to the model, which is illustrated in the simple Monte Carlo results displayed in Figure 1. But from a Bayesian perspective, the problem can be seen as the implicit use, in relying on OLS estimates, of a prior whose implications will in most applications be unreasonable.

The nature of the problem can be seen from analysis of the model that adds a constant term to (1),

$$y(t) = c + \rho y(t-1) + \varepsilon(t) \,. \tag{2}$$

This model can be reparameterized as

$$y(t) = (1-\rho)C + \rho y(t-1) + \varepsilon(t) \,. \tag{3}$$

If $|\rho| \neq 1$, the term C in (3) is the unconditional mean of y. When we confront any particular sample for $t = 1, \ldots, T$, this model will separate the sample's observed variation into two components: a deterministic component, predictable from data up to time $t = 0$, and an unpredictable component. The predictable component has the form

$$\bar{y}(t) = (y(0) - C)\rho^t + C \,. \tag{4}$$

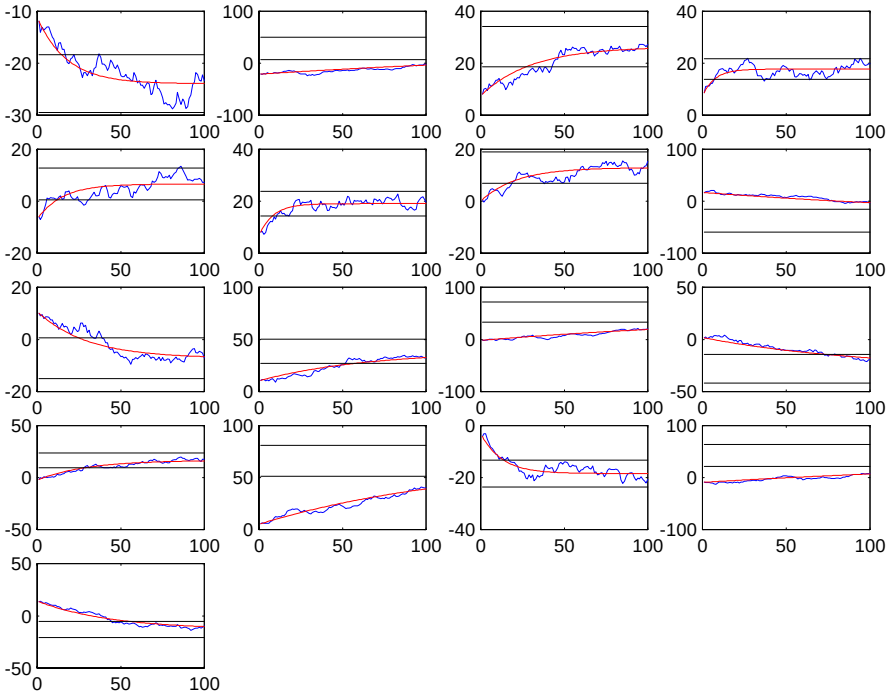

FIGURE 2. Initial Conditions Rogues' Gallery

Note: Rougher lines are Monte Carlo data. Smoother curved lines are deterministic components. Horizontal lines are 95% probability bands around the unconditional mean.

In the case where the true value of c is zero and the true value of ρ is one, the OLS estimates $(\hat{c}, \hat{\rho})$ of $(c, \rho) = (0, 1)$ are consistent. But $\hat{C} = \hat{c}/(1 - \hat{\rho})$ blows up as $T \to \infty$, because $\rho \to 1$ faster than $c \to 0$. This has the consequence that, even restricting ourselves to samples in which $|\hat{\rho}| < 1$, the proportion of sample variance attributed by the estimated model to the predictable component does not converge to zero, but rather to a nontrivial (and fat-tailed) limiting distribution. In other words, the ratio of the sum over the sample of $\bar{y}(t)^2$ (as computed from $(\hat{c}, \hat{\rho})$) to the sum over the sample of $y(t)^2$ does not converge to zero as $T \to \infty$, but instead tends to wander randomly.

Looking at Figure 2, we see plots of 16 extreme cases, out of 100 random draws generated from (2) with $y(0) \sim N(0, 100\sigma^2)$,$\rho = 1$, for the estimated amount of sample variance explained by the deterministic transient $\bar{y}(0)$. Note that in all of these cases the initial value of $y(0)$ is outside the two-standard-error band about C implied by the OLS estimates. Clearly the OLS fit is "explaining" linear trend and initial curvature observed in these sample paths as predictable from initial conditions.

The sample size in each plot in Figure 2 is 100, but the character of the plots is independent of sample size. This is a consequence of the scale invariance of Brownian motion. That is, it is well known that whether our sample of this discrete time random walk is of size 100 or 10,000, if we plot all the points on a graph of a given absolute width, and adjust the vertical scale to accommodate the observed range of variation in the data, the plot will look much the same, since its behavior will be to a close approximation that of a continuous time Brownian motion. Suppose we index the observations by their position $s = t/T$ on (0,1) when rescaled so the whole sample fits in the unit interval. Suppose further that we scale the data so that the final y in the scaled data $y^*(T) = y(T)\Big/\sqrt{T}$ has variance 1. If we reparameterize (2) by setting $\delta = 1/T$, $\gamma = c\delta$, $\phi = \rho - 1/\delta$, it becomes

$$y^*(s+\delta) - y^*(s) = \delta\phi \cdot (y^*(s) - C\sqrt{\delta}) + \sqrt{\delta}\varepsilon(t)\,. \tag{5}$$

Equation (5) is easily recognized as the standard discrete approximation,using small time interval δ, to the continuous time (Ohrnstein-Uhlenbeck) stochastic differential equation

$$dy^*(t) = \phi \cdot (y^*(s) - C\sqrt{\delta})\,dt + dW(t)\,. \tag{6}$$

But in the limit as $T \to \infty$, y^* is just a Brownian motion on (0,1). We know that we cannot consistently estimate the drift parameters of a stochastic differential equation from a finite span of data. So we expect that for large T we will have some limiting distribution for estimates of ϕ and $C\sqrt{\delta}$. The limiting distribution for estimates of ϕ implies a limiting distribution for estimates of $(1-\rho)/T = \phi$, which is a standard result. But what interests us here is that the portion of the sample variation attributed to a deterministic component will also tend to a non-zero limiting distribution.

Of course the situation is quite different, as $T \to \infty$, when the data come from (2) with $c \neq 0$ or $\rho \neq 1$. With $c \neq 0$ and $\rho = 1$, the data contain a linear trend component that dominates the variation. Our exercise of rescaling the data leads then, as $T \to \infty$, to data that looks just like a straight line, diagonally between corners of the graph. That is, in large samples the linear trend generated by $c \neq 0$ dominates the variation. Not surprisingly then, the position of the trend line on the scaled graph is well estimated in large samples in this case. If $|\rho| < 1$, then regardless of whether c is zero or not the estimated proportion of sample variance attributable to the deterministic component goes to zero as $T \to \infty$.

In applied work in economics, we do not know the values of c or ρ, of course, but we do often expect there to be a real possibility of c near enough to 0 and

ρ near enough to 1 that the kind of behavior of OLS estimates displayed in Figure 2, in which they attribute unrealistically large proportions of variation to deterministic components, is a concern.

It may be obvious why in practice this is "a concern", but to discuss the reasons formally we have to bring in prior beliefs and loss functions—consideration of what range of parameter values for the model are the center of interest in most economic applications, and what the consequences are of various kinds of error. Use of OLS estimates and the corresponding measures of uncertainty (standard errors, t-statistics, etc.) amounts to using a flat prior over (c, ρ). Such a prior is not flat over (C, ρ), but instead has density element $|1-\rho|\, dC\, d\rho$. That is, in (C, ρ) space it puts very low prior probability on the region where $\rho \doteq 1$. There is therefore an argument for simply premultiplying the usual conditional likelihood function by $1/|1-\rho|$. However, when as here different versions of a "flat prior" make a difference to inference, they are really only useful as a danger signal. The justification for using flat priors is either that they constitute a neutral reporting device or that, because often sample information dominates the prior, they may provide a good approximation to the posterior distribution for any reasonable prior. Here neither of these conditions is likely to hold. It would be a mistake to approach this problem by trying to refine the choice of flat prior, seeking a somehow "correct" representation of ignorance.

There may be applications in which the flat prior on (c, ρ) is a reasonable choice. We may believe that initial conditions at the start of the sample are unrepresentative of the steady state behavior of the model. The start date of the model might for example be known to be the date of a major historical break, like the end of a war. Barro and Sala-I-Martin (1995) (Chapter 11) display a number of cases where initial conditions are reasonably taken not to be drawn from the steady-state distribution of the process, so that a deterministic component is important.[1]

But we may instead believe that the sample period displays behavior that is representative of what we can expect over a period of similar length in the future. In that case we may be primarily interested in long term (meaning, say, over periods $T+1$ to $T+T$) forecasts for data after the end of the sample period. Results from samples like those displayed in the row and column positions (3,2), (3,4), (4,2) in Figure 2 are then particularly problematic. They show the initial conditions to be far outside a two-standard-error band about the estimated unconditional mean, but the terminal values $y(T)$ close to those

[1]Such cases lead to what they call "σ-convergence".

bands. Because of the nature of the exponential decay pattern imposed by the model, projections out of sample based on these estimates will show deterministic behavior qualitatively different from what has prevailed during the sample. The slope of the "trend line" will change, because the model implies that the data, over the time period $T+1, \ldots, 2T$, will be closer to its unconditional mean and thus less strongly subject to mean-reversion pressures. Of course this could be the way the future data will actually behave, but believing this requires a firm commitment to the restrictions on long-run behavior implicit in the parametric model. In accepting this prediction, we are agreeing to use a pattern in the observable data to predict a different pattern of variation, never yet observed, in the future data.

There can be no universally applicable formula for proceeding in this situation. One possibility, emphasized in an earlier paper of mine (Sims 1989), is that we believe that a wide range of complicated mechanisms generating predictable low frequency variation are possible. What looks like a positive linear trend might just be a rising segment of a sine wave with wavelength exceeding the sample size, or of a second or third order polynomial whose higher-order terms start having strong effect only for $t >> T$. In this case the proper course is to recognize this range of uncertainty in the parameterization. The result will be a model in which, despite a good fit to low frequency variation in the sample, uncertainty about long run forecasts will be estimated as high.

Another possibility, though, is that we believe we can use the existing sample to make long run forecasts, because we think it unlikely that there are such complicated components of variation at low frequencies. In this case, we believe that OLS estimates that look like most of those in Figure 2 are a priori implausible, and we need to use estimation methods that reflect this belief.

One way to accomplish this is to use priors favoring pure unit-root low-frequency behavior. An extreme version of this strategy is to work with differenced data. A less extreme version uses a reference prior or dummy observations to push more gently in this direction. The well-known Minnesota prior for vector autoregressions (Litterman 1983) (Doan 1992, Chapter 8) has this form, as does the prior suggested by Sims and Zha (1998) and the dummy observation approach in Sims (1993). A pure unit root generates forecasts that do not revert to a mean. It implies low-frequency behavior that will be similar in form within and outside the sample period. By pushing low-frequency variation toward pure unit root behavior, we tend to prevent it from showing oscillations of period close to T.

But this method has its limits. In large models, especially with high-frequency data (and correspondingly many lagged variables) it may require unreasonably strong emphasis on unit-root-like behavior in order to eliminate oscillatory low frequency deterministic components. Another possibility, not yet extensively tested in practice, would be to penalize unreasonable deterministic low frequency components directly in the prior. For example, one could include as a factor in the prior p.d.f. a p.d.f. for the ratio of the sample variance of the Fourier component at frequency $2\pi/T$ of $\Delta\bar{y}(t)$ to the variance of $\varepsilon(t)$.

There are open research questions here, and few well tested procedures known to work well in a wide variety of applications. More research is needed—but on how to formulate reasonable reference priors for these models, not on how to construct asymptotic theory for nested sequences of hypothesis tests that seem to allow us to avoid modeling uncertainty about low frequency components.

3. Dynamic, Possibly Non-Stationary, Panel Data Models

Here we consider models whose simplest form is

$$y_i(t) = c_i + \rho y_i(t-1) + \varepsilon_i(t)\ ,\ i = 1, \dots, N,\ t = 1, \dots, T. \tag{7}$$

Often in such models N is large relative to T. The common practice in time series inference of using as the likelihood the p.d.f. of the data conditional on initial conditions and on the constant term therefore runs into difficulty. Here there are N initial conditions $y_i(0)$, and N constant terms c_i, one of each for each i. The amount of data does not grow relative to the number of free parameters, so they cannot be estimated consistently as $N \to \infty$. This makes inference dependent, even in large samples, on the distribution of the c_i. For a Bayesian approach, the fact that the c_i have a distribution and cannot be estimated consistently raises no special difficulties. There is no distinction in this approach between random c_i and "parametric" c_i. In approaches based on pre-sample probability, by contrast, treating the c_i as random seems to require a different model ("random effects"), whose relation to the "fixed effect" model can seem a philosophical puzzle.[2]

[2]The classical distribution theory for the fixed effect model describes the randomness in estimators as repeated samples are generated by drawing ε vectors with the values of c_i held fixed. For the random effects models it describes randomness in estimators as repeated samples are generated by drawing ε vectors and c vectors. Though these two approaches usually are used to arrive at different estimators, each in fact implies a different distribution for any single estimator. That is, there is a random effects distribution theory for the fixed

If the dynamic model is known to be stationary ($|\rho| < 1$), with the initial conditions drawn randomly from the the unconditional distribution, then

$$y_i(0) \sim N\left(\frac{c_i}{1-\rho}, \frac{\sigma^2}{(1-\rho^2)}\right), \tag{8}$$

which we can combine with the usual p.d.f. conditional on all the $y_i(0)$'s to construct an unconditional p.d.f. for the sample. Maximum likelihood estimation with this likelihood function will be consistent as $N \to \infty$.

But if instead it is possible that $|\rho| \geq 1$, the distribution of $y_i(0)$ cannot be determined automatically by the parameters of the dynamic model. Yet the selection rule or historical mechanism that produces whatever distribution of $y_i(0)$ we actually face is critical to our inference. Once we have recognized the importance of application-specific thinking about the distribution of initial conditions for the non-stationary case, we must also acknowledge that it should affect inference for any case where a root R of the dynamic model may be expected to have $1/(1-R)$ of the same order of magnitude as T, which includes most economic applications with relatively short panels. When the dynamics of the model work so slowly that the effects of initial conditions do not die away within the sample period, there is generally good reason to doubt that the dynamic mechanism has been in place long enough, and uniformly enough across i, so that we can assume $y_i(0), i = 1, \ldots, N$ to be drawn randomly from the implied stationary unconditional distribution for $y_i(0)$.

It is well known that, because it makes the number of "parameters" increase with N, using the likelihood function conditional on all the initial conditions leads to bad results. This approach leads to MLE's that are OLS estimates of (7) as a stacked single equation containing the $N+1$ parameters $\{c_i, i = 1, \ldots, N, \rho\}$. These estimates are not consistent as $N \to \infty$ with T fixed. An alternative that is often suggested is to work with the differenced data. If $|\rho| \leq 1$, the distribution of $\Delta y_i(t), t = 1, ..., T$ is the same across i and a function only of σ^2 and ρ. Therefore maximum likelihood estimation based on the distribution of the differenced data is consistent under these assumptions. Lancaster (1997) points out that this approach emerges from Bayesian analysis under an assumption that $|\rho| < 1$, in the limit as the prior on the c_i becomes flat. (This Bayesian justification does not apply to the case where $|\rho| = 1$, however, where the method is nonetheless consistent.)

effects estimator and vice versa. There is no logical inconsistency in this situation, but if one makes the mistake of interpreting classical distribution theory as characterizing uncertainty about parameters given the data, the existence of two distributions for a single estimator of a single parameter appears paradoxical.

Using the p.d.f. of the differenced data amounts to deliberately ignoring sample information, however, which will reduce the accuracy of our inference unless we are sure the information being ignored is unrelated to the parameters we are trying to estimate.[3] If we are confident that $(c_i, y_i(0))$ pairs are drawn from a common distribution, independent across i and independent of the ε_i, then there is information in the initial y's that is wasted when we use differenced data.

A simple approach, which one might want to modify in practice to reflect application-specific prior information, is to postulate

$$(c_i, y_i(0)) \sim N\left(\begin{bmatrix}\mu_c \\ \mu_y\end{bmatrix}, \begin{bmatrix}\sigma_{cc} & \sigma_{cy} \\ \sigma_{cy} & \sigma_{yy}\end{bmatrix}\right), \tag{9}$$

with the parameters of this distribution treated as unknown. This unconditional joint c.d.f. can then be combined with the conventional conditional c.d.f. for $\{y_i(t), t = 1, \ldots, T,\ i = 1, \ldots, N\}$ conditional on the initial $y_i(0)$'s and c_i's, to produce a likelihood function. All the usual conditions to generate good asymptotic properties for maximum likelihood estimates are met here as $N \to \infty$.

Note that the setup in (9) is less restrictive than that in (8). The specification in (8) gives a distribution for $y_i(0)$ conditional on ρ, σ^2, and c_i. If we use it to generate a likelihood, we are implicitly assuming that none of these parameters affect the marginal distribution of c_i. If we complete the specification by postulating $c_i \sim N(\mu_c, \sigma_{cc})$, we have a special case of (9) in which

$$\frac{\sigma_{yc}}{\sigma_{cc}} = \frac{1}{1-\rho} \tag{10}$$

$$\sigma_{yy} - \frac{\sigma_{yc}^2}{\sigma_{cc}} = \frac{\sigma^2}{1-\rho^2}\,. \tag{11}$$

These two restrictions reduce the number of free parameters by two, but do not contradict the more general specification (9). The more general specification gives consistent estimates as $N \to \infty$ regardless of the size of $|\rho|$.

Of course where one is confident of the assumption of stationarity, it is better to use the restrictions embodied in (8), but in many economic applications it will be appealing to have a specification that does not prejudge the question of stationarity.

[3]Ignoring observed data by constructing a likelihood for a subset of or a function of the data generally gives us a different, less informative, likelihood function than we would get from all the data. But if the conditional distribution of the omitted data given the data retained does not depend on the parameters we are estimating, the likelihood for the reduced data set is the same as for the full data set.

No claim of great originality or wide usefulness is being made here for the specification in (9). Similar ideas have been put forth before. The panel data context, because it is the home of the random effects specification, has long been one where classically trained econometricians have had less inhibition about crossing the line between "parameters" and "random disturbances". Heckman (1981) proposes a very similar approach, and Keane and Wolpin (1997) a somewhat similar approach, both in the context of more complicated models where implementation is not as straightforward.

Lancaster (1997) has put forward a different Bayesian proposal. He looks for a way to reparameterize the model, redefining the individual effects so that a flat prior on them does not lead to inconsistent ML estimates. This leads him to different priors when likelihood is the unconditional p.d.f. for the data imposing stationarity and when likelihood is conditional on initial conditions. In the latter case he finds that this leads to an implicit prior on the c_i in the original parameterization that imposes a reasonable pattern of dependence between c_i and $y_i(0)$. His specification does enforce an absence of dependence between c_i and $y_i(0)$ when $\rho = 1$. While this will sometimes be reasonable, it is restrictive and may sometimes distort conclusions. For example, if we are studying growth of income $y_i(t)$ in a sample of countries i where the rich ones are rich because they have long had, and still have, high growth rates, the c_i and $y_i(0)$ in the sample will be positively correlated.

Applying the ideas in this section to more widely useful models in which right-hand-side variables other than $y_i(t-1)$ appear raises considerable complications. A straightforward approach is to collect all variables appearing in the model into a single endogenous variable vector y_i, and then re-interpret (7) and (9), making c_i a column and ρ a square matrix. Of course this converts what started as a single-equation modeling problem into a multiple-equation problem, but at least thinking about what such a complete system would look like is essential to any reasonable approach.

Another approach, closer to that suggested by Heckman (1981), preserves the single equation framework under an assumption of strict exogeneity. We generalize (7) to

$$y_i(t) = \rho y_i(t-1) + X_i(t)\beta_i + \varepsilon_i(t)\,, \tag{12}$$

in which we assume $X_i(t)$ independent of $\varepsilon_j(t)$ for all i, j, s, t in addition to the usual assumption that $\varepsilon(t)$ is independent of $y_j(s)$ for all $j \neq i$ with $s \leq t$ and for all $s < t$ with $j = i$. Now we need to consider possible dependence not just between a pair $(y_i(0), c_i)$, but among the vector $\{X_i(s), \beta_i, y_i(0), s = 1, \ldots, T\}$.

It is still possible here to follow basically the same strategy, though: postulate a joint normal unconditional distribution for this vector and combine it with the conditional distribution to form a likelihood function. One approach is to formulate the distribution as conditional on the X process, so that it takes the form

$$\begin{bmatrix} y_i(0) \\ \beta_i \end{bmatrix} \Bigg| \{X_i(s), s = 1, \ldots, T\} \sim N(\Gamma \mathbf{X_i}, \Sigma_{by}) \,, \tag{13}$$

where $\mathbf{X_i}$ is the T-row matrix formed by stacking $\{X_i(t),\, t = 1, \ldots, T\}$. Despite the fact that the model specification implies that $y_i(t)$ depends only on $X_i(s)$ for $s \leq t$, we need to allow for dependence of the initial y on future X's because they are related to the unobservable $X_i(s)$ for $s \leq 0$. If $N >> T$, this should be a feasible, albeit not easy, approach to estimation.

4. Kernel Estimates for Nonlinear Regression

Kernel methods for estimating nonlinear regressions are generally used when there is a large amount of data. They represent an attempt to represent a priori uncertainty more realistically than is possible in models with a small number of unknown parameters, and as a result they lead to implicitly or explicitly expanding the parameter space to the point where the degrees of freedom in the data and in the model are of similar orders of magnitude.

Kernel estimates of nonlinear regression functions do not emerge directly from any Bayesian approach. They do, though, emerge as close approximations to Bayesian methods under certain conditions. Those conditions, and the nature of the deviation between ordinary kernel methods and their Bayesian counterparts, show limitations and pitfalls of kernel methods. The Bayesian methods are interesting in their own right. They allow a distribution theory that connects assumptions about small sample distributions to uncertainty about results in particular finite samples, freeing the user from the arcane "order-$N^{q/p}$" kernel-method asymptotics whose implications in actual finite samples often seem mysterious.

The model we are considering here is

$$y_i = c + f(x_i) + \varepsilon_i \,, \tag{14}$$

where the ε_i are i.i.d. with variance σ^2, mean 0, independent of one another and of all values of x_j. We do not know the form of f, but have a hope of estimating it because we expect eventually to see a well dispersed set of x_i's and we believe that f satisfies some smoothness assumptions. A Bayesian formalization of these ideas makes f a zero-mean stochastic process on x-space

and c a random variable with a spread-out distribution. Though these ideas generalize, we consider here only the case of one-dimensional real $x_i \in \mathbb{R}$. Computations are simple if we postulate joint normality for c, the $f(\cdot)$ process and the ε's. It is natural to postulate that $c \sim N(0, \nu^2)$, with ν^2 large, and that f follows a Gaussian zero-mean stationary process, with the covariances of f's at different x's given by

$$\operatorname{cov}(f(x_1), f(x_2)) = R_f\left(|(x_1 - x_2)|\right) \tag{15}$$

for some autocovariance function R_f. We can vary the degree of smoothness we assume for f (including, for example, how many derivatives its sample paths have) by varying the form of R_f.

The distribution of the value $\bar{y}(x^*) = c + f(x^*)$ of the regression function $c + f$ at some arbitrary point x^*, given observations of $(y_i, x_i), i = 1, \ldots, N$, is then normal, with mean a linear function of the observed y's. To be explicit, (14) and (15) together imply

$$\operatorname{cov}\left(\{y_i, i = 1, \ldots, N\}\right) = \Omega = \nu^2 \underset{N\times N}{\mathbf{1}} + \underset{N\times N}{\left[R_f(x_i - x_j)\right]} + \sigma^2 I \tag{16}$$

and

$$\operatorname{cov}\left(\{y_i, i = 1, \ldots, N\}, \bar{y}(x^*)\right) = \Psi(x^*) = \nu^2 \underset{N\times 1}{\mathbf{1}} + \left[R_f(x_i - x^*)\right]_{i=1}^{N}, \tag{17}$$

where $\mathbf{1}$ is notation for a matrix filled with one's. The distribution of $\bar{y}(x^*)$ is then found by the usual application of projection formulas for Normal linear regression, yielding

$$E\left[\bar{y}(x^*)\right] = [y_1, \ldots, y_N] \cdot \Omega^{-1}\Psi(x^*) \tag{18}$$

$$\operatorname{var}\left[\bar{y}(x^*)\right] = R_f(0) - \Psi(x^*)'\Omega^{-1}\Psi(x^*)\,. \tag{19}$$

A kernel estimate with kernel k, on the other hand, estimates $\bar{y}(x^*)$ as

$$\hat{y}_k(x^*) = \frac{1}{N}\sum_{i=1}^{N} y_i k(x^* - x_i) = [y_1, \ldots, y_N] \cdot \left[k(x^* - x_i)\right]_{i=1}^{N}. \tag{20}$$

Comparing (18) with (20) we see that the Bayesian estimate is exactly a kernel estimate only under restrictive special assumptions. Generally, the presence of a non-diagonal Ω matrix in (18) makes the way an observation (y_i, x_i) is weighted in estimating $\bar{y}(x^*)$ depend not only on its distance from x^* (as in a kernel estimate), but also on how densely the region around x_i is populated with observations. This makes sense, and reflects how kernel methods are applied in practice. Often kernel estimates are truncated or otherwise modified

in sparsely populated regions of the x_i space, with the aim of avoiding spurious fluctuations in the estimates.

There are a variety of classical approaches to adapting the kernel to local regions of x space (Härdle 1990, section 5.3). However, these methods emphasize mainly adapting bandwidth locally to the estimated shape of f, whereas the Bayesian approach changes the shape of the implicit kernel as well as its bandwidth, and does so in response to data density, not the estimated shape of f. A Bayesian method that adapted to local properties of f would emerge if $\text{cov}(f(x_i), f(x_j))$ were itself modeled as a stochastic process.

There are conditions, though, under which the Bayesian estimates are well approximated as kernel estimates, even when Ω is far from being diagonal. If the x_i values are equally spaced and sorted in ascending order, then Ω is a Toeplitz form (i.e. constant down diagonals), and to a good approximation, at least for i not near 1 or N, we can write (18) as

$$E\left[\bar{y}(x^*)\right] = \hat{y}_k(x^*) \tag{21}$$

for

$$k = N \cdot \left((\sigma^2\delta + R_f + \nu^2)^{-1} * (R_f + \nu^2) - b\right) , \tag{22}$$

where b is a constant term that adjusts the integral of the kernel to 1.[4] In (22) δ is the δ-function, equal to zero except when its argument is zero, the "$*$" refers to convolution[5] and the inverse is an inverse under convolution. Thus kernel estimates are justifiable as close to posterior means when the data are evenly dispersed and we are estimating $\bar{y}(x^*)$ for x^* not near the boundary of the x-space. These formulas allow us to see how beliefs about signal-to-noise ratio (the relative size of R_f and σ^2) and about the smoothness of the f we are estimating are connected to the choice of kernel shape. Note that for $\sigma^2 >> R_f$, we have $k \doteq N\sigma^{-2}R_f - b$, so that the kernel shape matches R_f. However it is not generally true that the kernel shape mimics that of the R_f that characterizes our prior beliefs about smoothness.

Samples in economics often do show unevenly dispersed x values. Many applications to cross section data involve x's like firm size or individual income, which have distributions with notoriously fat tails, for example. Bayesian methods allow us to account explicitly, and automatically, for variations in the density of sampling across x-space. Understanding of their connection to

[4]The effect of allowing for a large-prior-variance c in (14) is concentrated on making the integral of the kernel emerge as one.

[5]In discrete time, the convolution of a function h with a function g is the new function $h * g(t) = \sum_s h(s)g(t-s)$.

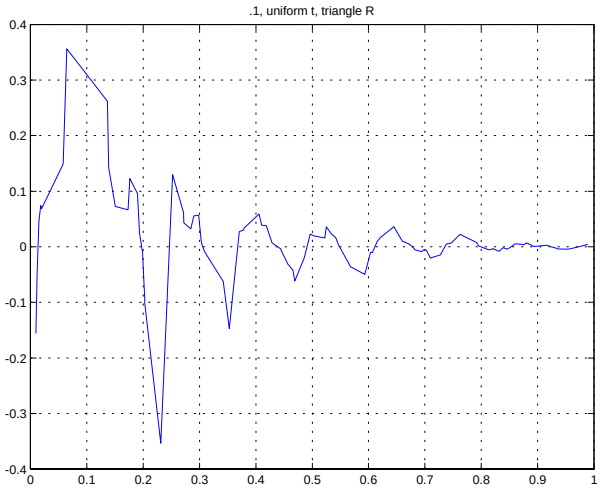

FIGURE 3. Weighting function for $x^* = .1$
Nearest x_i: .0579, .0648, .1365, .1389

kernel estimation methods gives insight into why it generally makes sense to adapt kernel shape to the density of sampling in a given region—if only by making special adjustments at the boundaries of the x-space.

To illustrate these points we show how they apply to a simple example, in which $\{x_i,\ i = 1, \ldots, 100\}$ are a random sample from a uniform distribution on $(0, 1)$ and our model for f makes $R_f(t) = \max(1 - |t| / .12, 0)$. We assume $\sigma = .17$, making the observational error quite small relative to the variation in f, and set $\nu = 10$.[6] In Figure 3 we see weights that reflect the fact that $x^* = .1$ happens to fall in the biggest gap in the x_i data in this random sample. The gap is apparent in the list of 4 adjacent x_i's given below the figure. In contrast, Figure 4 shows a case where the sample happens to contain many x_i's extremely close to x^*. In this case there is no need to rely on distant x_i's. Simply averaging the ones that happen to fall nearly on top of x^* is the best option. The implied kernel also tends to stretch out at the boundaries of the sample, as is illustrated in Figure 5, which is for x^*=1.

The Bayesian approach to nonlinear regression described here is not likely to replace other approaches. If R_f is chosen purely with an eye toward reflecting reasonable prior beliefs about the shape of f, the result can be burdensome computations when the sample size is large. The Ω matrix is of order $N \times N$, and if it has no special structure the computation of $\Omega^{-1}\Psi(x^*)$ is much more

[6]The effects of ν on the implicit kernels we plot below are essentially invisible. The effects would not be invisible for estimated f values if the true c happened to be large.

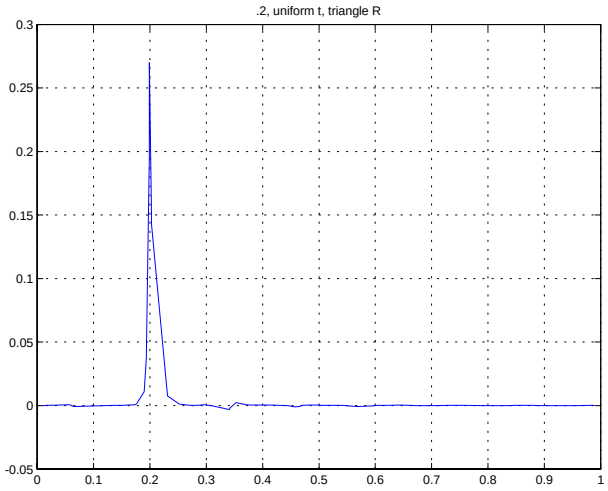

FIGURE 4. Weighting function for $x^* = .2$
Nearest x_i: 6 between .1934 and 0.2028

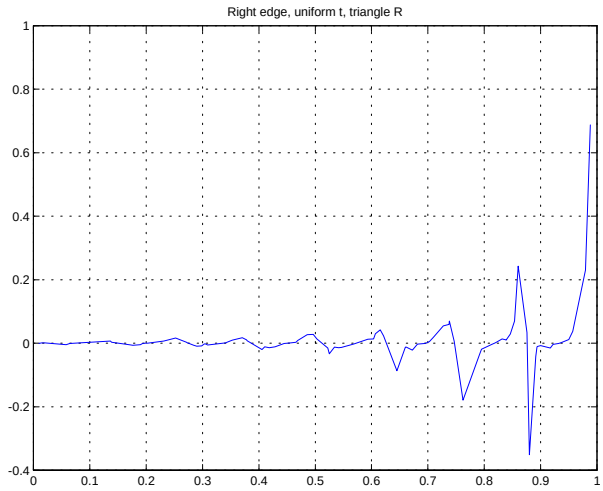

FIGURE 5. Weighting function for $x^* = 1$
Nearest x_i: .9501 .9568 .9797 .9883

work than simply applying an arbitrarily chosen kernel function. The amount of work can be held down by, as in the example we have discussed, keeping the support of R_f bounded, which makes Ω a constant matrix plus a matrix whose non-zero elements are concentrated near the diagonal. It can be held down even further by special assumptions on R_f. Grace Wahba, in a sustained program of theoretical and applied research (Wahba 1990, Luo and Wahba 1997), has developed "smoothing spline" methods. In their specialization to

the type of problem discussed in this section, they can be derived from the assumption that the p'th derivative ($p \geq 0$) of the f process is a Brownian motion. In the example we have considered, our assumptions imply

$$f(x) = W(x) - W(x - .12)\,, \tag{23}$$

where W is a Brownian motion. Thus the special computational methods for smoothing splines do not apply to the example, though it can easily be verified that the estimated f in the example will be a continuous sequence of linear line segments, and thus a spline.[7] Generally, the result of the Bayesian method is a spline whenever R_f is a spline, though it will not be a smoothing spline *per se* for any choice of a stationary model for f.

5. Conclusion

The examples we have considered are all situations where realistic inference must confront the fact that the complexity of our ignorance is of at least the same order of magnitude as the information available in the data. Assumptions and beliefs from outside the data unavoidably shape our inferences in such situations. Formalizing the process by which these assumptions and beliefs enter inference is essential to allowing them to be discussed objectively in scientific discourse. In these examples, approaching inference from a likelihood or Bayesian perspective suggests some new interpretations or methods, but even where it only gives us new ways to think about standard methods such an approach will be likely to increase the objectivity and practical usefulness of our analysis by bringing our probability statements more into line with common sense.

References

Barro, R. J., and X. Sala-I-Martin (1995): *Economic Growth*. McGraw-Hill, New York.

Doan, T. A. (1992): "RATS User's Manual, Version 4," Estima, Evanston, IL.

Härdle, W. (1990): *Applied Nonparametric Regression*, vol. 19 of *Econometric Society Monographs*. Cambridge University Press, Cambridge, New York, New Rochelle, Melbourne, Sydney.

Heckman, J. J. (1981): "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time–Discrete Data Process," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C. F. Manski, and D. McFadden, pp. 179–195, Cambridge, Massachusetts. MIT Press.

[7] This follows from the fact that $\Psi(x^*)$ is by construction linear in x^* over intervals containing no values of x_i or of $x_i \pm .12$.

Hurwicz, L. (1950): "Least Squares Bias in Time Series," in *Statistical Inference in Dynamic Economic Models*, ed. by T. C. Koopmans, Cowles Commission Monograph Number 10, New York. Wiley.

Keane, M. P., and K. I. Wolpin (1997): "Career Decisions of young Men," *Journal of Political Economy*, 105, 473–522.

Lancaster, T. (1997): "Orthogonal Parameters and Panel Data," Working Paper 97-32, Brown University.

Litterman, R. B. (1983): "A random walk, Markov model for the distribution of time series," *Journal of Business and Economic Statistics*, 1, 169–73.

Luo, Z., and G. Wahba (1997): "Hybrid Adaptive Splines," *Journal of the American Statistical Association*, 92, 107–116.

Sims, C. A. (1989): "Modeling Trends," in *Proceedings*, American Statistical Association Annual Meetings.

——— (1993): "A 9 Variable Probabilistic Macroeconomic Forecasting Model," in *Business Cycles, Indicators, and Forecasting*, ed. by J. H. Stock, and M. W. Watson, vol. 28 of *NBER Studies in Business Cycles*, pp. 179–214.

Sims, C. A., and H. D. Uhlig (1991): "Understanding Unit Rooters: A Helicopter Tour," *Econometrica*.

Sims, C. A., and T. Zha (1998): "Bayesian Methods for Dynamic Multivariate Models," *International Economic Review*.

Wahba, G. (1990): *Spline Models for Observational Data*, vol. 59 of *CBMS-NSF Regional Conference Series in Applied Mathematics*. Society for Industrial and Applied Mathematics, Philadelphia.