THINKING ABOUT INSTRUMENTAL VARIABLES

CHRISTOPHER A. SIMS

ABSTRACT. We take a decision-theoretic view on the question of how to use instrumental variables and method of moments. Since prior beliefs play an inevitably strong role when instruments are possibly "weak", or when the number of instruments is large relative to the number of observations, it is important in these cases to report characteristics of the likelihood beyond the usual IV or ML estimates and their asymptotic (i.e. second-order local) approximate standard errors. IV and GMM appeal because of their legitimate claim to be convenient to compute in many cases, and a (spurious) claim that they can be justified with few "assumptions". We discuss some approaches to making such a claim more legitimately.

I. INTRODUCTION

Situations where we apply IV and GMM are good illustrations of the value of a Bayesian perspective, because in these cases, in contrast to most situations, in large samples Bayesian analysis treating the likelihood as the posterior (flat-prior analysis) does not turn out to be always equivalent to treating frequentist pre-sample probability statements as if they were valid post-sample probability statements.

A Bayesian perspective on scientific reporting views all data analysis as reporting of the shape of the likelihood in ways likely to be useful to the readers of the report. We examine moment-based inference from that perspective.

II. THE LITERATURE

If you want a through discussion of the mechanics of Bayesian analysis of instrumental variables, there is much recent progress, with particularly good discussions in Geweke (1996) and Kleibergen and Zivot (2000). There is also a recent line of work, to which Kitamura and Stutzer (1997), Zellner, Tobias, and Ryu (1997), and Kim (2002) among others have contributed, that attempts to "derive" IV, in some cases with a post-sample posterior accompanying it, by using automatic, information-theory-based methods to avoid explicit priors. Related work by Phillips and Chao pursues the use of Jeffreys priors, which are another automatically generated class of priors.

Date: March 6, 2007.

Key words and phrases. Bayesian, GMM, Instrumental Variables, Weak Instruments, Instrument Selection, entropy.

The work on Bayesian mechanics develops insight into the unusual aspects of the likelihood in models with simultanaeity that arise from the poles and zeros in the mapping between reduced form and structural parameters. Part of this paper develops examples that try to show the importance of these peculiarities in practice. The work on automatic priors and posteriors is motivated by recognizing the appeal of the GMM and IV claim that they require few assumptions, while at the same time being dissatisfied with that claim as a guide to practice. We begin this paper with a discussion of weak assumptions.

III. Assumption avoidance

Whatever assumptions one makes are likely to be disputable, so it is comforting to an econometrician facing a critical professional audience to be able to claim that his conclusions depend on hardly any assumptions. Models that seem to make few assumptions are often characterized as "nonparametric", in contrast to assumption-laden parametric models. Some years ago when frequency-domain time series methods first appeared in econometrics, they were initially characterized as allowing time series inference with much weaker assumptions than the parametric ARIMA models that were, and still are, common in econometrics. I pointed out in a survey paper then (1974) that this claim was spurious. Frequency domain theory obtains *asymptotic* sampling theory with few assumptions, but it does so by proposing to smooth frequency-domain statistics through windows or kernels that will narrow in width as sample size increases and by making smoothness assumptions on the frequency-domain objects that are the objects of interest. Everyone understands that there are functions in the class of objects of interest that are allowed by the weak assumptions used to derive asymptotic theory, but that in fact have to be disallowed if we are take seriously assertions about uncertainty (confidence intervals, tests of hypotheses, etc.) in the sample at hand. Whatever the kernel bandwidth, there will be "smooth" functions that oscillate too wildly to be accurately estimated by a kernel of that bandwidth. We intuitively apply these restrictions in evaluating results, even though they often are not discussed in research reports. With specific application to hypothesis testing, these are the issues discussed in Faust (1999).

In the time series area it fairly quickly became widely understood that there was actually no increased generality in kernel-based frequency-domain methods. ARIMA models in which the number of parameters is allowed to grow systematically with sample size or as a function of the data can generate an asymptotic distribution theory fully as general as the frequency-domain theory. The class of time series models that can be well approximated by a sequence of parametric models is no larger, in a topological sense, than the class of models that satisfy the smoothness assumptions underlying frequency-domain methods. Smoothness of functions in function spaces is the same kind of assumption, topologically, as rate-of-decay of sequences in sequence spaces.

These issues reappear in the same form in the theory of non-parametric estimation of regression functions or density functions. My impression, though, is that practitioners outside the time series area are less aware of the fact that explicitly nonparametric methods are only a different guess at what kind of model will work well in a particular application, not a way of obtaining results with fewer assumptions.[1]

The issues also appear in the standard claims that inference based on OLS, IV, and GMM is free of assumptions on distributions of disturbances. These claims in practice are used to support statements about uncertainty of results that, in a given model and sample, are justfiable only by restricting the class of disturbance distributions more tightly than is required in the asymptotic theory. In many cases, the distribution theory justified by the asymptotics would be exactly correct under normality assumptions on the residuals, and correct to a reasonable approximation in a given sample only if the disturbances are within some neighborhood of normality.

In effect, we end up justifying normality assumptions in inference by appealing to asymptotics. This suggests that the assumption has a chance of being robust. Entropy-based reasoning suggests the assumption may also be in a certain sense conservative, however.

IV. The Model

$$\underset{T\times 1}{y} = x\beta + \varepsilon = \underset{T\times k}{Z}\gamma\beta + \nu\beta + \varepsilon \tag{1}$$

$$\underset{T\times 1}{x} = Z\gamma + \nu \tag{2}$$

$$\mathrm{Var}([\nu\beta + \varepsilon \;\; \nu]) = \Sigma \tag{3}$$

Note that the same model can be parameterized "in reverse", without changing the implied class of possible probability models for the data, by simply switching the positions of y and x in these equations. This symmetry is brought out further if we write $Y = [y\ x]$, $\Pi = [\gamma\beta\ \gamma]$, $\eta = [\nu\beta + \varepsilon \;\; \nu]$ and then the model as

$$Y = Z\Pi + \eta\,. \tag{4}$$

The model then requires that the $k \times 2$ matrix Π be of rank 1, and once we have imposed that restriction, this last form of the model (4) again traces out the same set of probability models for the data.

[1]Probably this is not, or at least not only, because time series econometricians are smarter. Frequency domain models in time series are very awkward if one needs to project forward in time from data available at that date. Since this is a standard task in applied time series analysis, it was a relief to practitioners to have theorists prove that non-parametric models are not inherently more general. In nonparametric regression, on the other hand, the local character of the smoothness restrictions implicit in kernel methods is often more reasonable than the restrictions implied by a standard, say orthogonal polynomial, parametric model. Practitioners are therefore less in need of excuses not to use the kernel methods.

When we use the parametrization (1-2), likelihood does not go to zero as $\beta \to \infty$ with Σ fixed, no matter what the sample size. This is because for very large β, the best fit will be obtained with very small $\|\gamma\|$, but with γ chosen so that $\beta\gamma$ gives the best possible fit to the y equation (1). This will of course make the fit of the x equation (2) poor in most cases, but in the limit of very large β, very small $\|\gamma\|$, the x equation will simply have approximately zero on the right-hand side. Thus the fit deteriorates as $\beta \to \infty$, but only toward a definite limiting value that does not push likelihood to zero.

This creates pitfalls for naive Bayesian approaches. Markov chain Monte Carlo (MCMC) methods applied directly to the likelihood as if it were an unnormalized pdf will not converge, and analytically integrating parameters out of the likelihood may fail, or may produce improper marginal "posteriors".

A flat prior on Π in (4), with no restriction on Π's rank (the unrestricted reduced form, or URF), in contrast does produce an integrable posterior if the sample size is not extremely small. It therefore seems promising to use Euclidean distance to define a metric on the reduced-rank submanifold of Π, then transform the flat prior (Lebesgue measure) on this submanifold to β, γ coordinates. This derivation does not in itself actually guarantee that posteriors under this improper prior will be proper, but it is a promising approach. The improper prior on β, γ that emerges from this approach is

$$\left| \frac{\partial \Pi}{\partial(\beta,\gamma)} \left(\frac{\partial \Pi}{\partial(\beta,\gamma)} \right)' \right|^{\frac{1}{2}} = \|\gamma\| \, (1+\beta^2)^{\frac{1}{2}} \, . \tag{5}$$

It does lead to proper posteriors.

Even with this prior, however, the posteriors decline only at a polynomial rate in the tails, and the degree of the polynomial does not increase with sample size. This is in contrast with the posteriors under a flat prior in a linear regression model, where the tails decline at a polynomial rate that increases linearly with sample size.

Consider now what this means when we combine a prior that has Gaussian tails or tails that decrease as a high-order polynomial with the likelihood weighted by (5). If we start with the prior mean and the likelihood peak lined up with each other, then let the likelihood peak move away from the prior mean while keeping the shapes of the likelihood and the prior pdf fixed, the posterior mean, median and mode move away from the prior mean (as expected) at first, *but then reverse direction, coming back to coincide with the prior mean when the likelihood peak is very far from the prior mean.* This is illustrated graphically in Figure 1.

In other words, in an instrumental variables model, the fat tails of the likelihood imply that when the likelihood is sufficiently in conflict with the prior, the prior dominates, even though the likelihood may have a sharp enough peak to dominate prior information that is not so far from the likelihood peak. This has important practical implications for interpreting IV estimates — which is indeed reflected in the widespread recognition that there is a "weak instrument" problem. But from the Bayesian perspective, weak instruments is not a characteristic of a

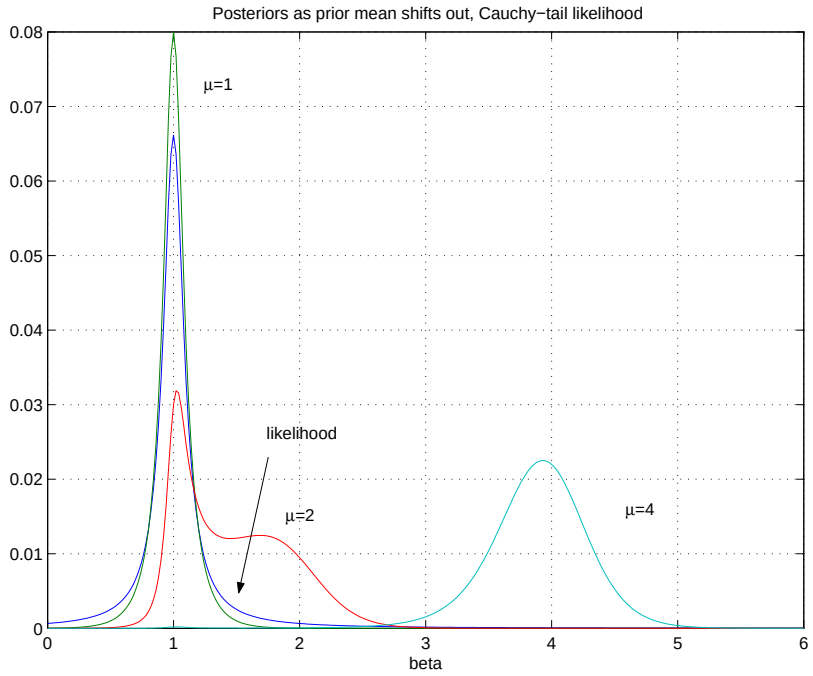

FIGURE 1
Posterior as distance of MLE from prior mean increases. Prior is t with 9 d.f., $\sigma = 1/3$. Likelihood is Cauchy, $\sigma = .1$, center 1.

population, or even of a sample, but rather of the interaction of a sample and prior beliefs. IV estimators occasionally produce "erratic" estimators, and practitioners understand that. There is no way to "test" for whether such an estimate is erratic, however; one has to consider prior beliefs, and whether the sample likelihood in the neighborhood of reasonable values for β is indeed nearly logarithmically flat.

V. JOINT POSTERIORS, MARGINALS FOR THE JID CASE

Figures 2 and 3 show the contours of the posterior for the just identified case, where the joint posterior is easily computed analytically. Note the strongly non-elliptical shapes, which imply that prior information on γ will have strong implications for the shape and precision of the posterior distribution of β, and vice versa. Also note that the plot for the case of a prior flat on β and γ shows quite complicated behavior near $\|\gamma\| = 0$.

The marginal posterior on β implied by Figure 3 is displayed in Figure 4. Note how misleading the local Gaussian approximation is in this case.

VI. WEAK INSTRUMENT CASE

In a sample where the posterior is highly non-Gaussian and has substantial, slowly declining tails, even apparently very weak prior information can substantially affect inference. The graphs displayed here show the effects of imposing a $N(0, 100I)$ prior on β, γ jointly in a model and sample that imply an estimated β around 1 in magnitude, with substantial uncertainty. Even this weak prior has a

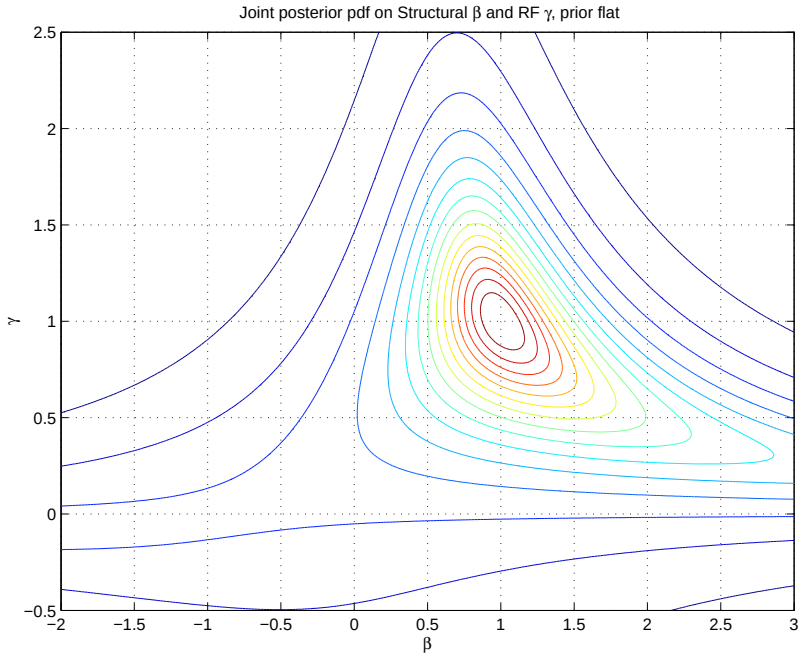

FIGURE 2

$$Z'Z = 4\,,\ \hat{\Sigma} = \begin{bmatrix} .67 & .33 \\ .33 & .67 \end{bmatrix}\,,\ \hat{\beta} = \hat{\gamma} = 1$$

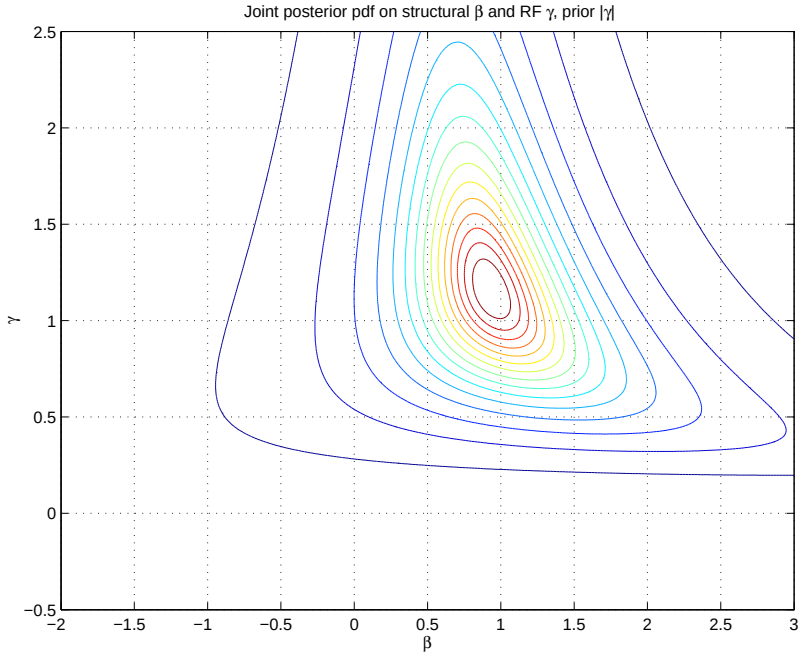

FIGURE 3
Notes: See figure 2

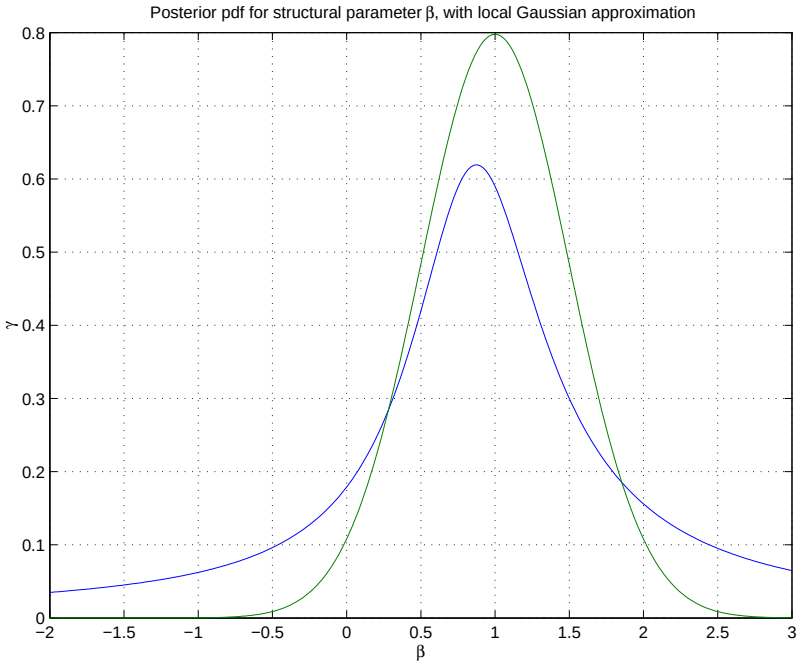

FIGURE 4
Notes: See Figure 2

dramatic effect on the posterior, largely by eliminating extremely spread-out tails. Certainly posterior means would be greatly affected by including such weak prior information.

To prepare these graphs I generated MCMC artificial samples of size 10,000, sampling successively from the conditional likelihoods of $\{\beta \mid \gamma, \Sigma\}$, $\{\gamma \mid \beta, \Sigma\}$, and $\{\Sigma \mid \gamma, \beta\}$, which are all of standard forms, then applying a Metropolis-Hastings step to reflect the influence of the prior. When the prior is not used, very large values of β occur in the sampling, and when β gets so large, dependence of β on $\|\gamma\|$ is very strong. As is well known, heavy dependence produces slow convergence of Gibbs samplers. Figures 10 and 11 illustrate how the use of the prior improves the convergence properties of the Gibbs sampler, by eliminating the extremely large draws of β.

VII. MANY-INSTRUMENT CASE

When $T \leq k$, i.e. when there are no degrees of freedom in the "first stage regression", the likelihood has two infinite peaks: One where γ is chosen to fit x perfectly in (2), the other where $\beta\gamma$ is chosen to fit y perfectly in (1). We usually don't find convincing the corresponding estimates, which amount to using OLS to estimate a regression of y on x or to estimate a regression of x on y, estimating β as the inverse of the coefficient in the reversed regression. This is because we don't actually believe that either of the two equations in the URF is likely to fit perfectly. It makes sense, then, to try to reflect this belief in a prior that might pull us away from the OLS estimates.

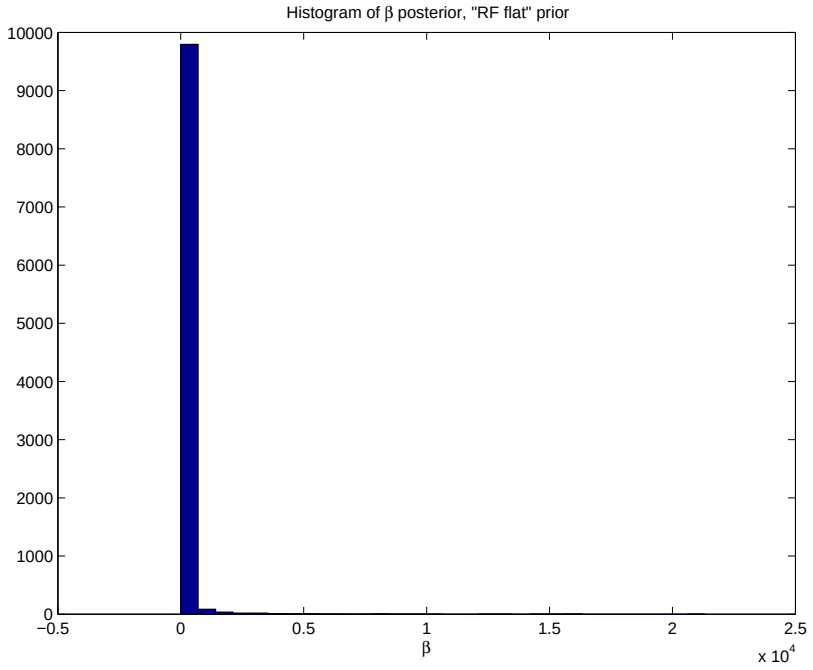

FIGURE 5

$$x = Z\begin{bmatrix}1\\.1\end{bmatrix} + \varepsilon, \quad y = Z\begin{bmatrix}1\\.1\end{bmatrix} + u$$
$$\underset{20\times 2}{Z}, \varepsilon, \nu \text{ all } iidN(0,1).$$

Actual IV $\hat{\beta} = 1.5132$ with asymptotic standard error 0.8412 .

One way to do this is to use a conjugate prior, which can be implemented by adding "dummy observations" to the data set. To be specific we take $T = k = 20$ and specify the expanded data as

$$x^* = \begin{bmatrix} x \\ \underset{T\times 1}{0} \end{bmatrix}, \quad y^* = \begin{bmatrix} y \\ \underset{T\times 1}{0} \end{bmatrix}, \quad Z^* = \begin{bmatrix} Z \\ \lambda \underset{T\times T}{I} \end{bmatrix}. \tag{6}$$

The calculations underlying Figure 12 used $\lambda = .5$. It is natural to interpret these dummy observations as tending to "shrink" the results toward zero. However, because they tie the prior variance of γ and $\beta\gamma$ about zero to the variance of the disturbance terms in the equations, they also downweight "perfect fits". It can be seen from Figure 12 that the latter effect dominates, as applying the prior has shifted the posterior upward, away from the prior mean (and from the OLS estimate). This approach seems to show some promise as a way to avoid throwing away information by not using clearly valid and available instrumental variables.

REFERENCES

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

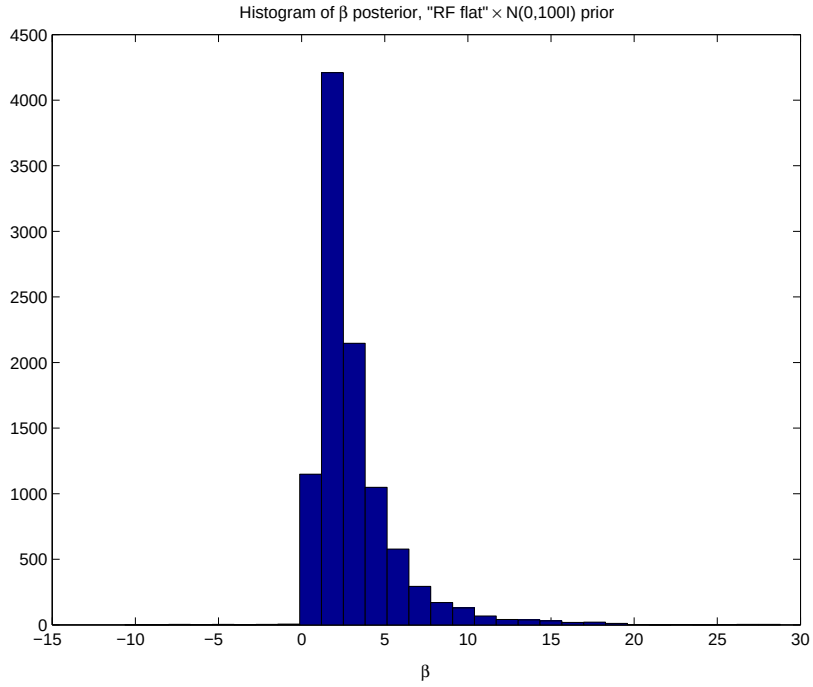

FIGURE 6
Note: Model and sample as in Figure 5

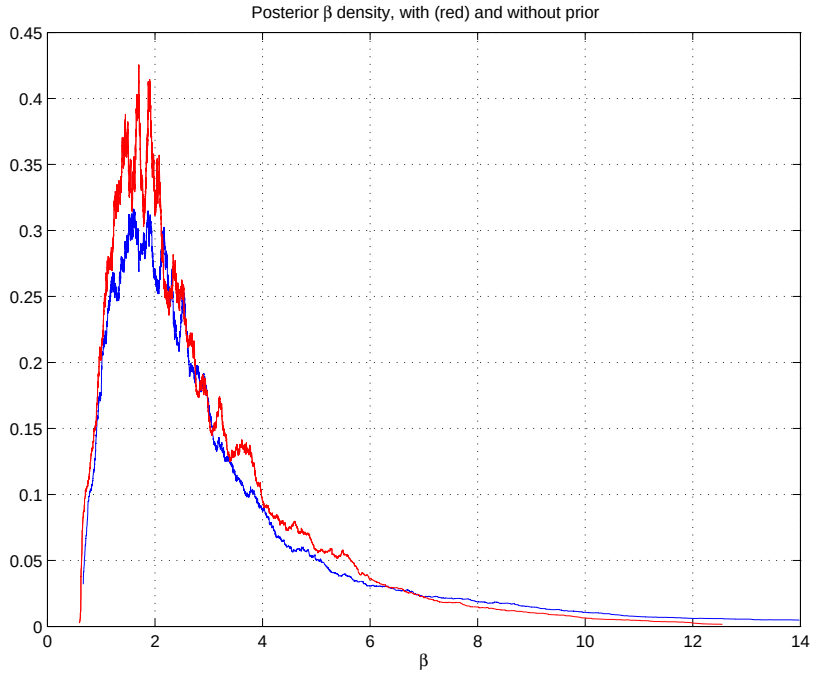

FIGURE 7
Note: Model and sample as in Figure 5. Densities estimated from a "300 nearest neighbor" calculation.

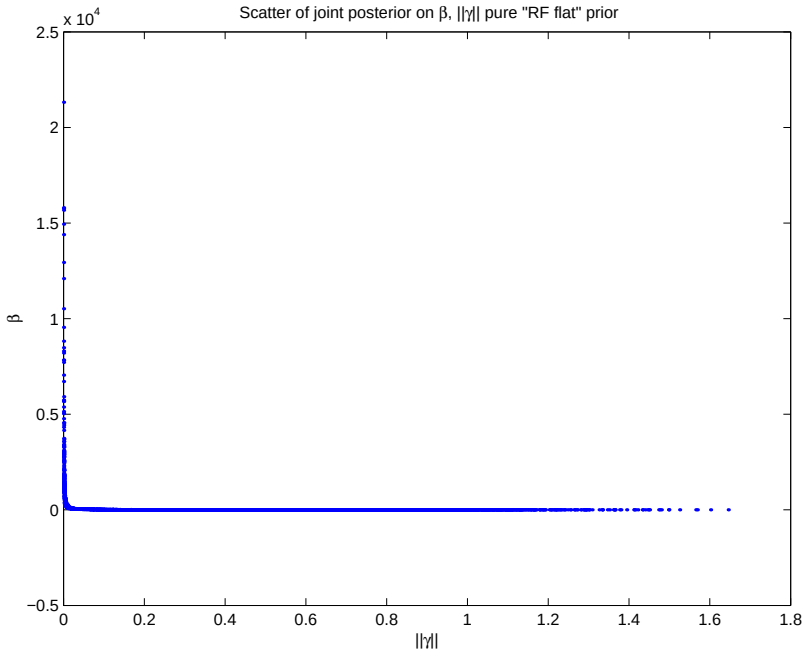

FIGURE 8
Note: Model and sample as in Figure 5

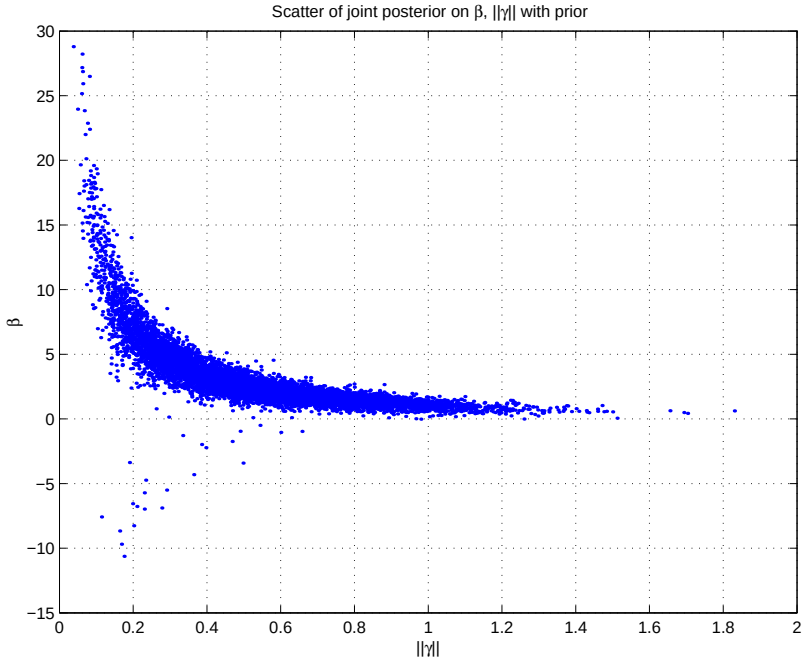

FIGURE 9
Note: Model and sample as in Figure 5

GEWEKE, J. (1996): "Bayesian Reduced Rank Regression in Econometrics," *Journal of Econometrics*, 75, 121–146.

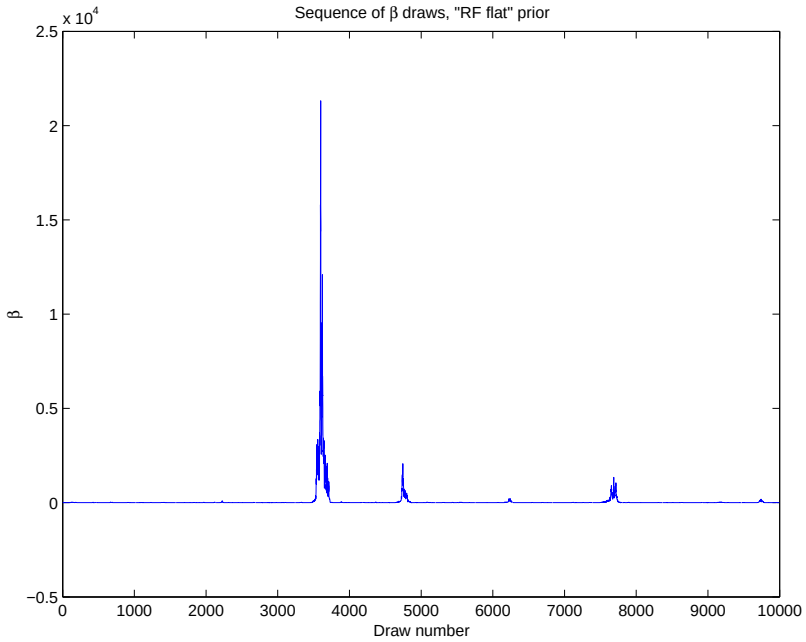

FIGURE 10
Note: Model and sample as in Figure 5

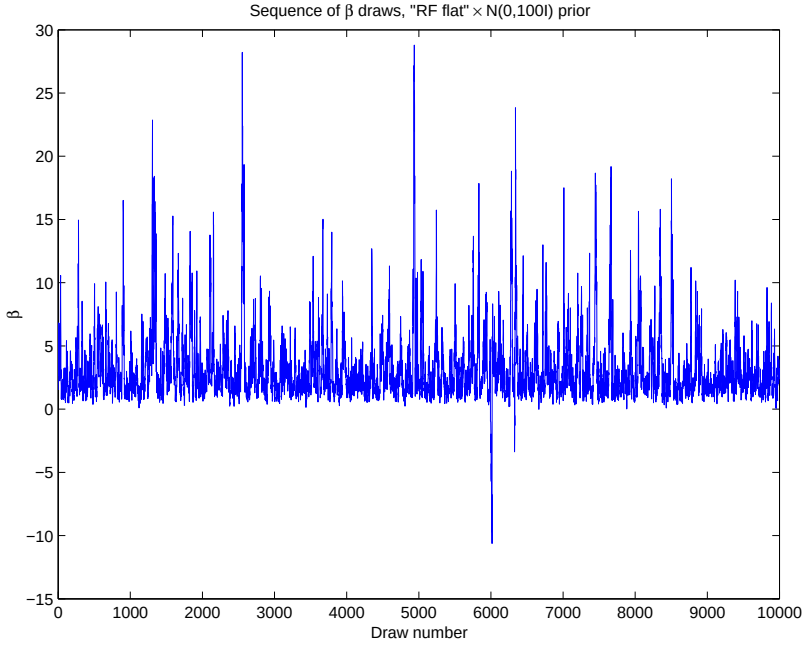

FIGURE 11
Note: Model and sample as in Figure 5

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

KITAMURA, Y., AND M. STUTZER (1997): "An Information-Theoretic Alternative to Generalized Method of Moments Estimation," *Econometrica*, 65(4), 861–874.

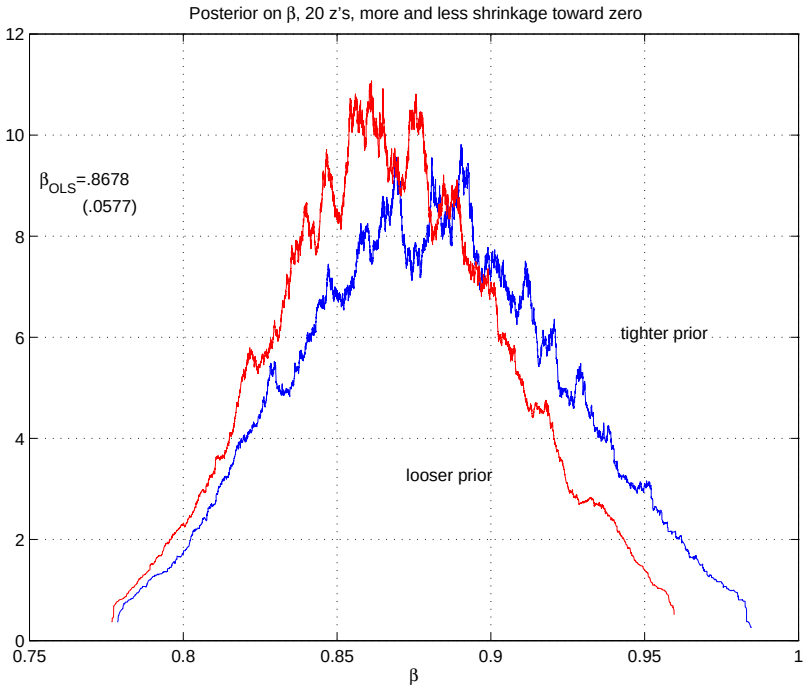

FIGURE 12
Z is 20×20, iid N(0,1). Coefficients on Z all 1 in reduced form. Sample has $\hat{\beta}_{OLS} = .8678$ with usual standard error estimate .0577. $\hat{\beta}$ from reversed regression is .9446.

KLEIBERGEN, F., AND E. ZIVOT (2000): "Bayesian and Classical Approaches to Instrumental Variable Regression," Discussion paper, Econometric Institute, Rotterdam.

SIMS, C. A. (1974): "Distributed Lags," in *Frontiers of Quantitative Economics II*, ed. by M. Intrilligator, and D. Kendrick. North-Holland.

ZELLNER, A., J. TOBIAS, AND H. K. RYU (1997): "Bayesian Method of Moments (BMOM) Analysis of Parametric and Semiparametric Regression Models," Discussion paper, University of Chicago.

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