A Bayesian perspective on GMM and IV

Christopher A. Sims
Princeton University
sims@princeton.edu

November 26, 2013

What is a Bayesian perspective?

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.

What is special about GMM and IV?

What is special about GMM and IV?

1. They are not derived from likelihood.

What is special about GMM and IV?

1. They are not derived from likelihood.

2. They are used mostly with only asymptotic theory to guides statements about uncertainty of inference.

What is special about GMM and IV?

1. They are not derived from likelihood.

2. They are used mostly with only asymptotic theory to guides statements about uncertainty of inference.

3. They are appealing because the assumptions they force us to make explicit — the moment conditions — often are intuitively appealing or even emerge directly from a respectable theory, while they allow us to sweep under the asymptotic rug “auxiliarly” assumptions.

What is special about GMM and IV?

1. They are not derived from likelihood.

2. They are used mostly with only asymptotic theory to guides statements about uncertainty of inference.

3. They are appealing because the assumptions they force us to make explicit — the moment conditions — often are intuitively appealing or even emerge directly from a respectable theory, while they allow us to sweep under the asymptotic rug "auxiliarly" assumptions.

4. They lead to easy computations.

Which of these are actually pluses?

- 1 and 2 are defects, and some Bayesians have taken the view that IV and GMM methods are just mistakes, because their lack of likelihood foundation and small sample distribution theory makes Bayesian interpretation of them seem difficult.

Which of these are actually pluses?

- 1 and 2 are defects, and some Bayesians have taken the view that IV and GMM methods are just mistakes, because their lack of likelihood foundation and small sample distribution theory makes Bayesian interpretation of them seem difficult.

- But 4, computational convenience, is certainly an advantage and a reason that Bayesians should not dismiss IV and GMM.

Which of these are actually pluses?

- 1 and 2 are defects, and some Bayesians have taken the view that IV and GMM methods are just mistakes, because their lack of likelihood foundation and small sample distribution theory makes Bayesian interpretation of them seem difficult.

- But 4, computational convenience, is certainly an advantage and a reason that Bayesians should not dismiss IV and GMM.

- And 3, while problematic, reflects a legitimate desire for inference that is not sensitive to assumptions that we sense are somewhat arbitrary.

Do asymptotics free us from assumptions?

Do asymptotics free us from assumptions?

- The short answer: no.

Do asymptotics free us from assumptions?

- The short answer: no.
- The usual Gaussian asymptotics can be given either a Bayesian or a non-Bayesian (pre-sample) interpretation. The Bayesian interpretation is that asymptotics gives us approximations to likelihood shape. (Kwan, 1998)

Do asymptotics free us from assumptions?

- The short answer: no.

- The usual Gaussian asymptotics can be given either a Bayesian or a non-Bayesian (pre-sample) interpretation. The Bayesian interpretation is that asymptotics gives us approximations to likelihood shape. (Kwan, 1998)

- If we wish to characterize the implications of a particular sample, we may decide that asymptotic theory is likely to be a good guide, or we may not. This is a decision-theoretic judgment call. We know, usually, that there are conditions on the true model that would imply that asymptotics are a good guide for this sample. Actually using the asymptotic theory amounts to judging, without explicit discussion, that it is ok to ***assume*** these conditions are met.

Examples of the “freedom from assumptions” fallacy: kernel methods

- Frequency domain methods in time series, kernel methods for regression. Frequency domain methods are “non-parametric” in comparison to ARMA models in the same mathematical sense that kernel methods are non-parametric in comparison to parametric polynomial regressions.

Examples of the “freedom from assumptions” fallacy: kernel methods

- Frequency domain methods in time series, kernel methods for regression. Frequency domain methods are “non-parametric” in comparison to ARMA models in the same mathematical sense that kernel methods are non-parametric in comparison to parametric polynomial regressions.

- In time series, after an initial burst of enthusiasm, the fact that there is no true increased generality in use of frequency-domain methods sank in (possibly in part because for any application involving forecasting, the FD methods are inconvenient). In cross-section non-parametrics, one still finds econometricians who think that kernel methods require “fewer assumptions”.

- Using kernel methods in a particular sample with a particular model obviously requires that the true spectral density or regression function not be badly distorted by convolution with the kernel. This greatly limits the class of admissible spectral densities or regression functions, in comparison with the class allowed by the asymptotic theory.

- Using kernel methods in a particular sample with a particular model obviously requires that the true spectral density or regression function not be badly distorted by convolution with the kernel. This greatly limits the class of admissible spectral densities or regression functions, in comparison with the class allowed by the asymptotic theory.

- There is a well defined topological sense in which it can be shown that a countable, dense class of finitely parameterized models is as "large" as the classes of functions that are allowed by the smoothness restrictions required for kernel-method asymptotics.

IV and GMM as assumption-free

- They make assertions about the distributions of only certain functions of the data — not enough functions of the data to fully characterize even the first and second moments.

IV and GMM as assumption-free

- They make assertions about the distributions of only certain functions of the data — not enough functions of the data to fully characterize even the first and second moments.

- They limit their assertions to a few moments, hence do not require a complete description of the distribution of even those functions of the data about which they do make assertions.

Embedding IV and GMM in a framework for inference

- Required: An *explicit* set of assumptions that will let us understand what we are implicitly assuming in applying IV or GMM in a particular sample, and thereby also let us construct likelihood or posterior distributions.

Embedding IV and GMM in a framework for inference

- Required: An *explicit* set of assumptions that will let us understand what we are implicitly assuming in applying IV or GMM in a particular sample, and thereby also let us construct likelihood or posterior distributions.

- Approach 1: Find assumptions that make the asymptotic theory exact. This may involve using the sample moments on which IV or GMM are based in somewhat different ways.

Embedding IV and GMM in a framework for inference

- Required: An *explicit* set of assumptions that will let us understand what we are implicitly assuming in applying IV or GMM in a particular sample, and thereby also let us construct likelihood or posterior distributions.

- Approach 1: Find assumptions that make the asymptotic theory exact. This may involve using the sample moments on which IV or GMM are based in somewhat different ways.

- Approach 2: Choose from among the models that are consistent with the asymptotic theory, a model that is “conservative”, either in the sense that it is robust (a good approximation to a large class of models) or that it draws the weakest possible inferences from the data, given the explicit assumptions.

Examples of these approaches: Approach 1

- LIML: Normality assumptions on disturbances imply both a likelihood function and asymptotic theory that matches that of IV.

- An analogue to LIML for GMM? If GMM is based on $E[g(y_t, \beta) \mid z_t] = 0$, then in the LIML case we are providing a (linear) model, not dependent on β, for the distribution of $\partial g / \partial \beta \mid z_t$. For GMM, it is not obvious that a linear model for this object is appropriate, and there are apparently many possible choices. There may be work on this issue of which I am not aware, but it seems ripe for exploration.

Examples continued: Approach 2

Empirical likelihood: Treat the joint distribution of data and parameters as concentrating all probability on the observed data points and as generated from a uniform distribution over all probabilities on those points that satisfy the moment restrictions. The usual procedure is then to find the mode of this distribution jointly in the probabilities of the data points and the parameters. But one can also take a Bayesian perspective and integrate over the probabilities to get a marginal on the parameters. Of course the assumption of probabilities concentrated on the observed points would be a terrible approximation for certain purposes.

Zellner BMOM: assume it is exactly true that $E[Z'u \mid y] = 0$, where this is the sample cross-moment matrix for observed instrument matrix Z and unobserved error vector u in the current sample, with expectation over the distribution of the parameter given the data. Then find the maximal entropy distribution for $\{\beta \mid \text{data}\}$ under this assumption. This exactly justifies the usual IV asymptotics and thus is an example of both approach 2 and approach 1. Of course the usual models out of which IV arises do not justify the basic assumption made to generate BMOM, except as an asymptotic approximation.

Kitamura and Stutzer maximal entropy GMM: Treat the joint distribution of data and parameters as concentrated on the observed data points. Derive the maximal entropy distribution of the parameters given the data under this assumption plus the moment restrictions. Closely related to empirical likelihood. More clearly "conservative", but harder to give a complete Bayesian interpretation: It generates a posterior, but it's not clear what joint distribution of data and parameters would lead to this posterior via Bayes' rule.

Can we do better?

- Both entropy-maximization and searching for exact theory that supports the use of asymptotics are useful ideas. Even exact theory (like BMOM) that requires strange assumptions is helpful in make it clear what we are implicitly assuming when we use the usual asymptotics.

Can we do better?

- Both entropy-maximization and searching for exact theory that supports the use of asymptotics are useful ideas. Even exact theory (like BMOM) that requires strange assumptions is helpful in make it clear what we are implicitly assuming when we use the usual asymptotics.

- But the BMOM assumption and the "probability concentrated on the observed points" assumptions are obviously unattractive.

Can we do better?

- Both entropy-maximization and searching for exact theory that supports the use of asymptotics are useful ideas. Even exact theory (like BMOM) that requires strange assumptions is helpful in make it clear what we are implicitly assuming when we use the usual asymptotics.

- But the BMOM assumption and the "probability concentrated on the observed points" assumptions are obviously unattractive.

- Gaussianity assumptions do emerge from maximizing entropy subject to first and second moment restrictions, which suggests that where normality assumptions lead to tractable likelihoods consistent with the asymptotics, normality may be a conservative assumption.

- However entropy is a slippery concept. It is not invariant to transformations of the random variable. More reliable, in my view, is the notion of **mutual information** between random variables. This is the expected reduction in entropy of the distribution of one random variable after we observe another random variable. It is invariant to transformations, which is another way of saying it depends only on the copula of the two jointly distributed variables.

- However entropy is a slippery concept. It is not invariant to transformations of the random variable. More reliable, in my view, is the notion of **mutual information** between random variables. This is the expected reduction in entropy of the distribution of one random variable after we observe another random variable. It is invariant to transformations, which is another way of saying it depends only on the copula of the two jointly distributed variables.

- We can imagine deriving conservative models that minimize the mutual information between data and parameters subject to moment restrictions. I have explored this idea a bit. It is clear that it will only work with proper priors, and that it will not deliver exactly standard GMM except as a limiting case.

Applying these ideas

- IV and GMM have two widely recognized breakdown regions: “Weak” instruments, and “overabundant” instruments. We mistrust the implications of conventional asymptotic inference in these situations, but non-Bayesian procedures give us little guidance in characterizing our uncertainty.

- Likelihood description here provides better guidance about the nature of uncertainty.

A simple model

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

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

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

or

$$y = z\theta + \xi \tag{4}$$

$$x = z\gamma + \nu \tag{5}$$

$$\theta = \gamma\beta\,. \tag{6}$$

- When we use the parametrization (1-2), likelihood does not go to zero as $\beta \to \infty$, no matter what the sample size.

- This can create problems for naive Bayesian approaches — MCMC not converging, integrating out nuisance parameters creating paradoxes.

- A flat prior on (θ, γ) with no rank restrictions 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 (θ, γ), 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

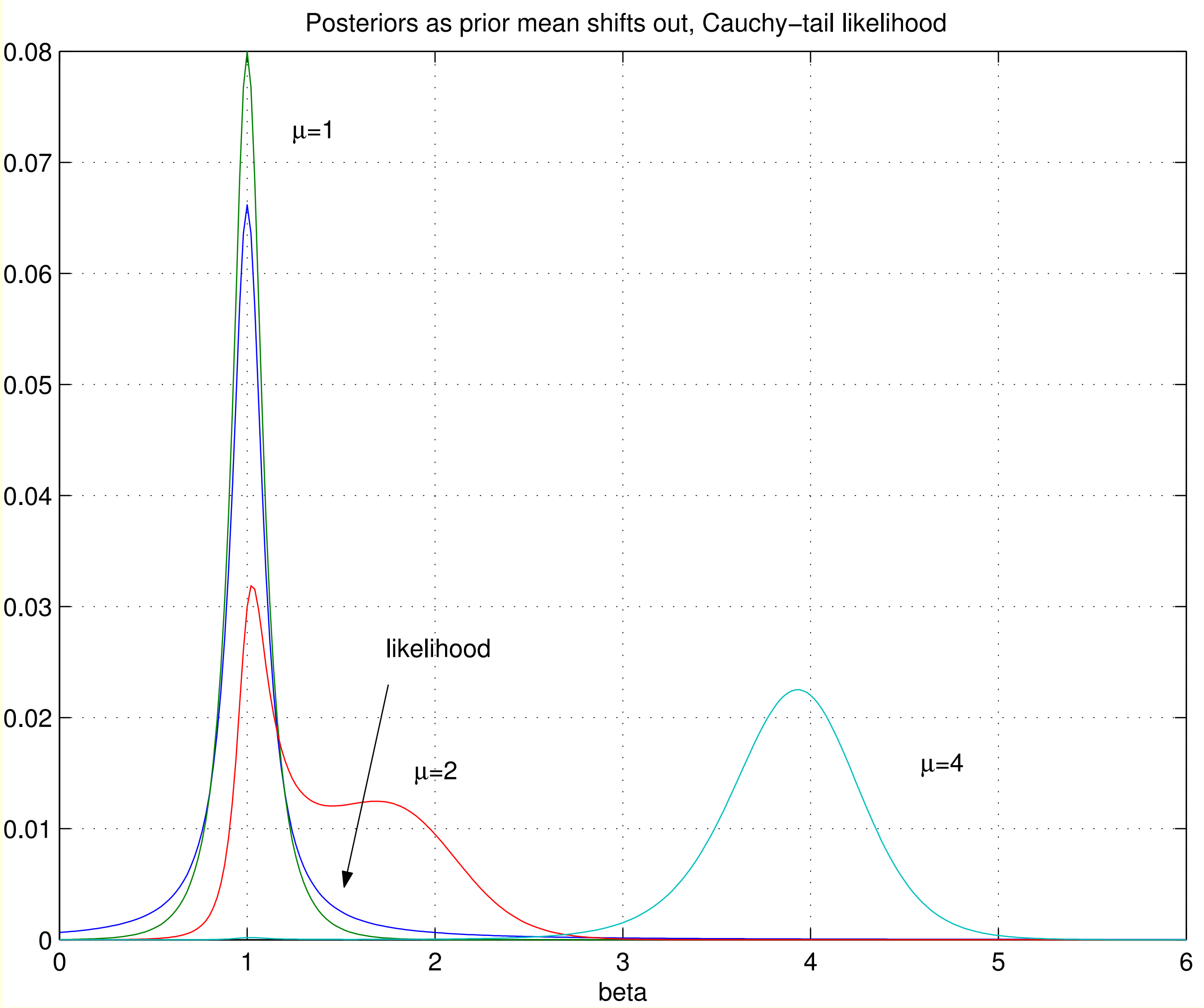
$$\left|\frac{\partial \Pi}{\partial(\beta,\gamma)}\left(\frac{\partial \Pi}{\partial(\beta,\gamma)}\right)'\right|^{\frac{1}{2}} = \|\gamma\|\,(1+\beta^2)^{\frac{1}{2}}\,. \tag{7}$$

  It does lead to proper posteriors.

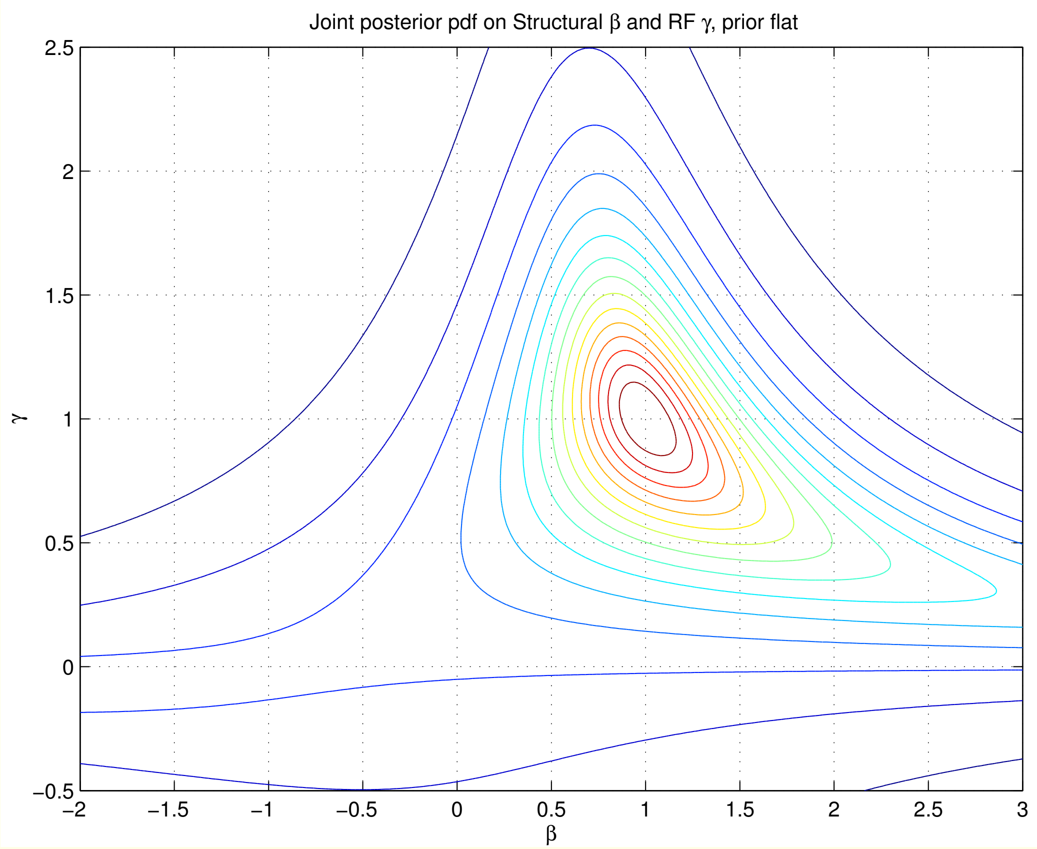
- Even with this prior, however, the posteriors decline only at a polynomial rate in the tails, and the degree of the polynomial does not increase with sample size. This is in contrast with the posteriors under a flat prior in a linear regression model, where the tails decline at a polynomial rate that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

Boomerang IV

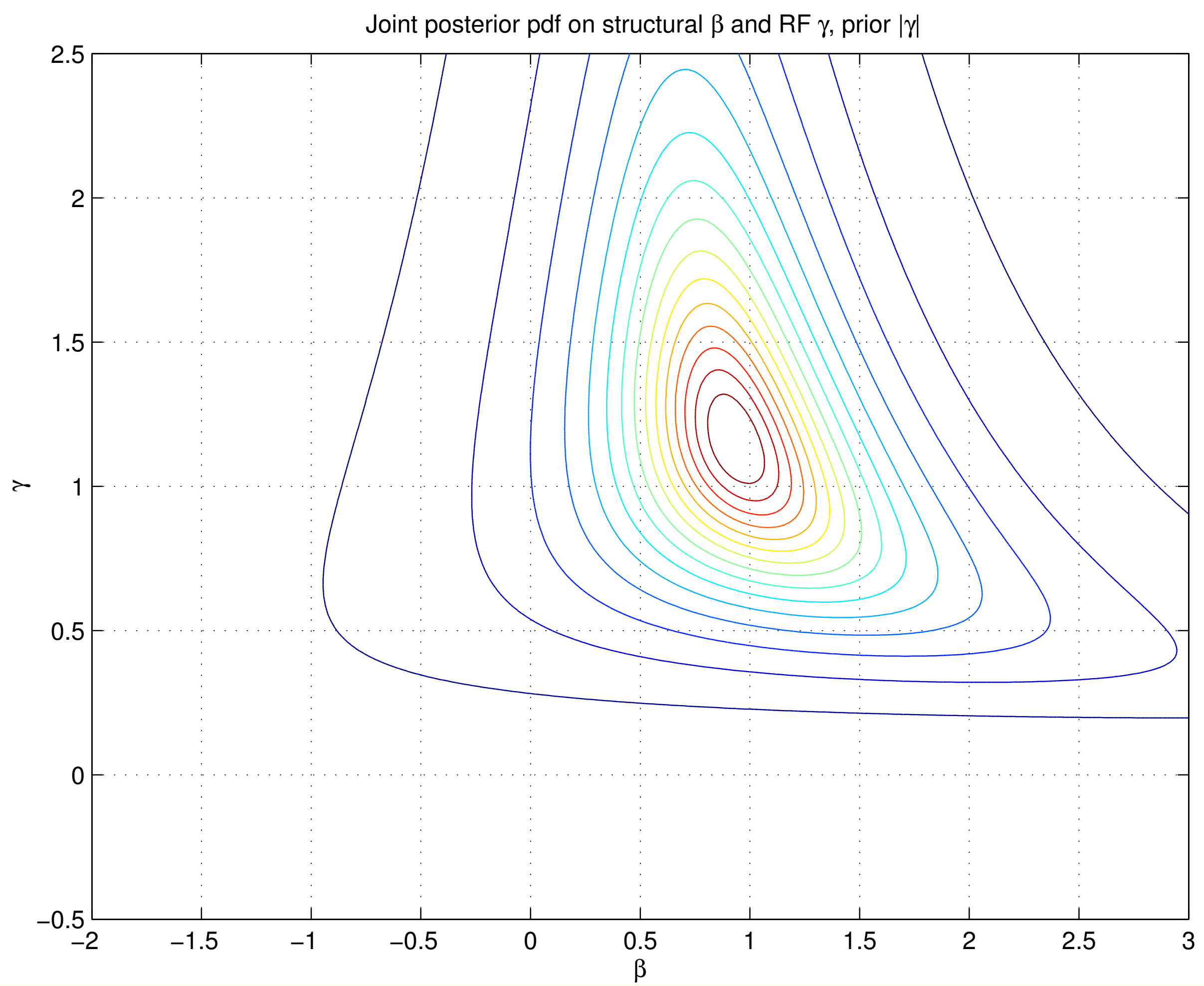
Consider now what this means when we combine a prior that has Gaussian tails or tails that decrease as a high-order polynomial with the likelihood weighted by (7). If we start with the prior mean and the likelihood peak lined up with each other, then let the likelihood peak move away from the prior mean 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 the figure.



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.

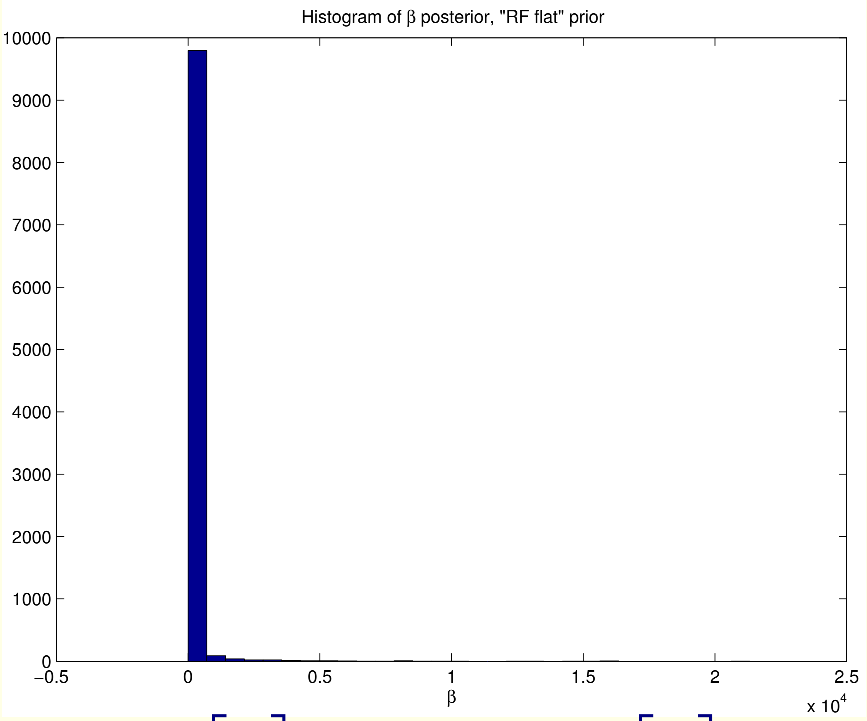


$$Z'Z = 4\,,\ \hat{\Sigma} = \begin{bmatrix} .67 & .33 \\ .33 & .67 \end{bmatrix}\,,\ \hat{\beta} = \hat{\gamma} = 1$$


Joint posterior pdf on structural β and RF γ, prior |γ|
2.5
2
1.5
1
0.5
0
−0.5
γ
−2
−1.5
−1
−0.5
0
0.5
1
1.5
2
2.5
3
β

Weak instrument case

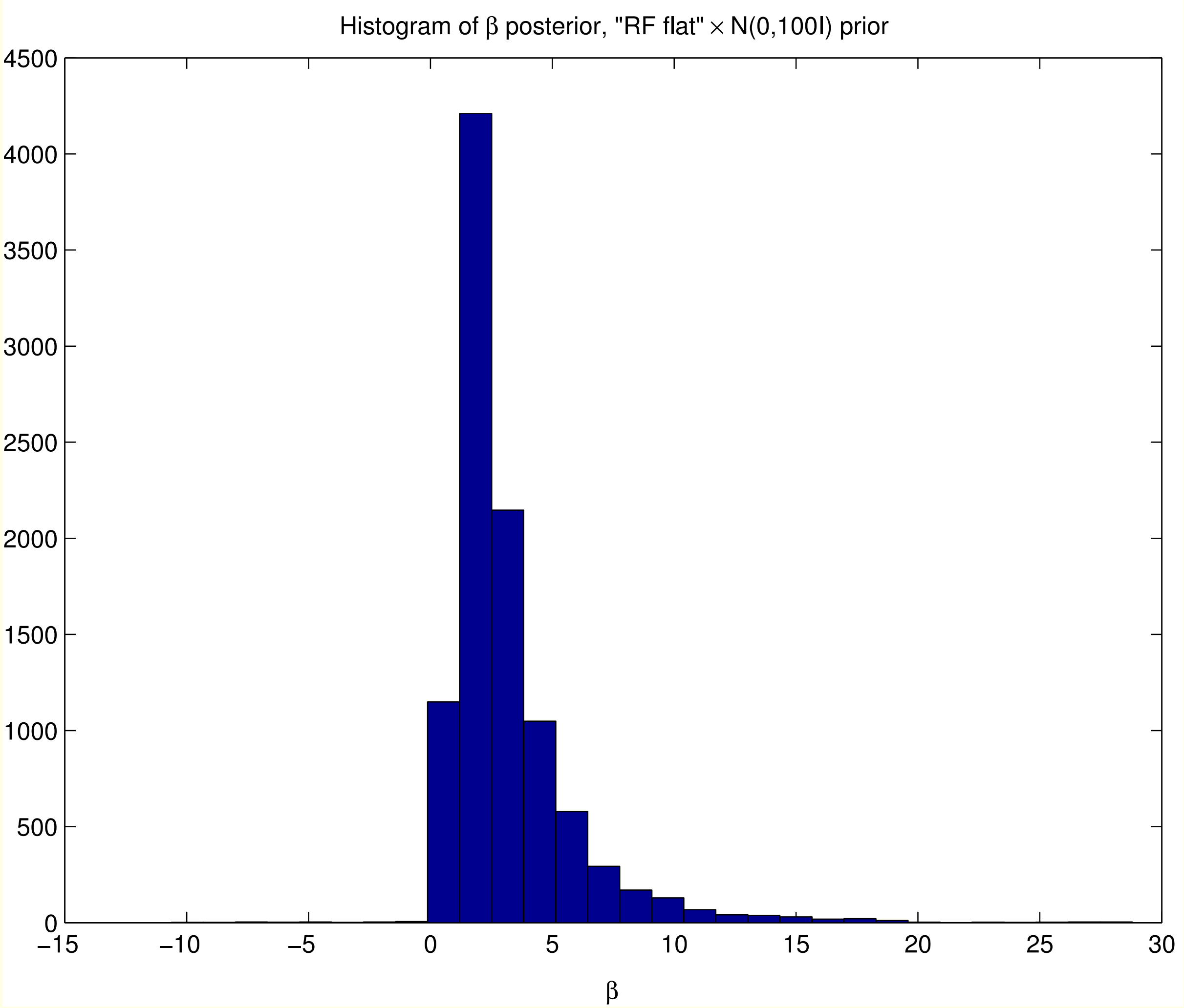
In a sample where the posterior is highly non-Gaussian and has substantial, slowly declining tails, even apparently very weak prior information can substantially affect inference. The graphs displayed here show the effects of imposing a $N(0, 100I)$ prior on β, γ jointly in a model and sample that imply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

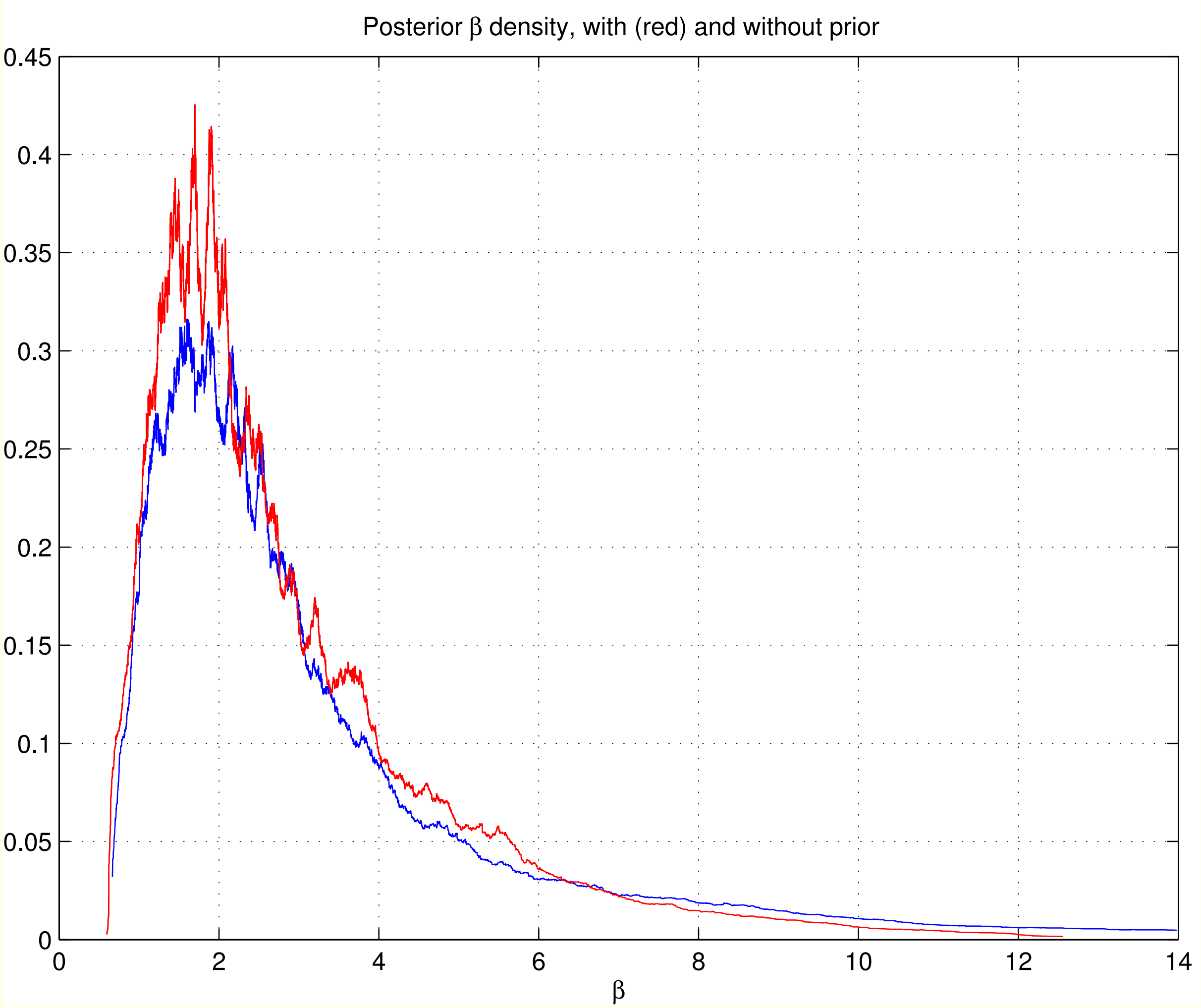


$$x = Z \begin{bmatrix} 1 \\ .1 \end{bmatrix} + \varepsilon, \quad y = Z \begin{bmatrix} 1 \\ .1 \end{bmatrix} + u$$

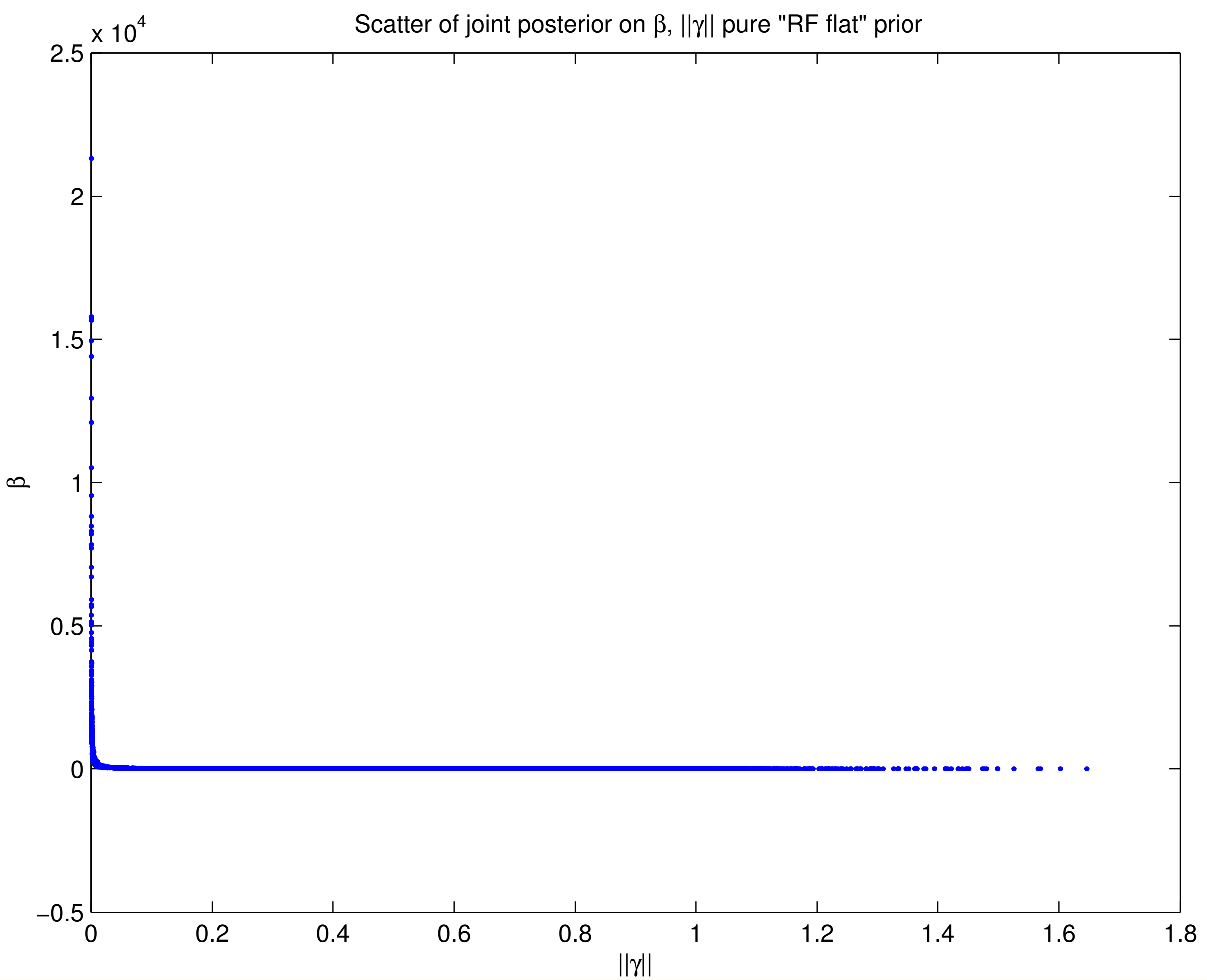
$\underset{20\times 2}{Z}, \varepsilon, \nu$ all $iidN(0,1)$.

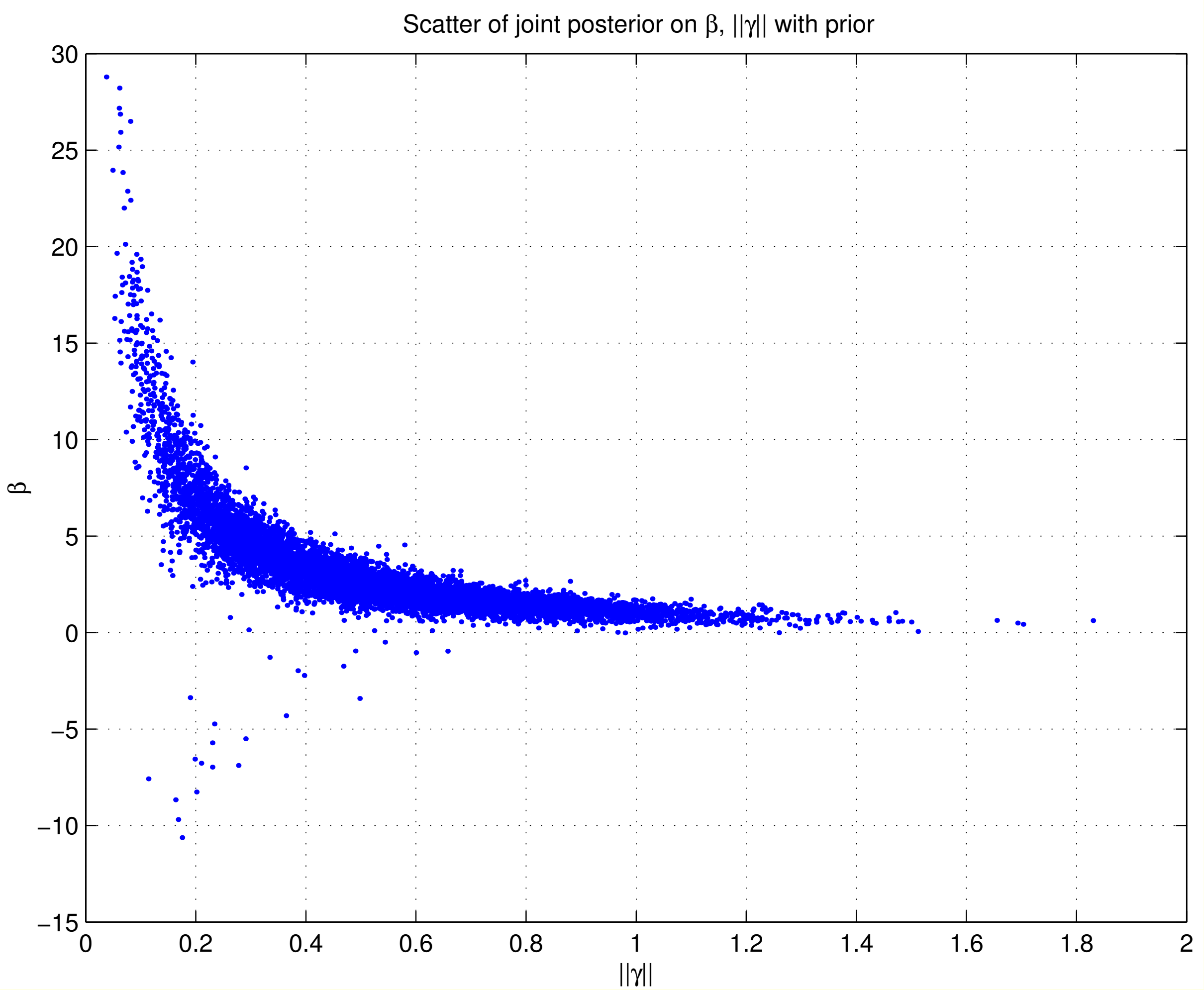
Actual IV $\hat{\beta} = 1.5132$ with asymptotic standard error 0.8412


Histogram of β posterior, "RF flat" × N(0,100I) prior
4500
4000
3500
3000
2500
2000
1500
1000
500
0
–15
–10
–5
0
5
10
15
20
25
30
β



Note: Densities estimated from a “300 nearest neighbor” calculation.


Scatter of joint posterior on β, ||γ|| pure "RF flat" prior
x 10^4
2.5
2
1.5
1
0.5
0
−0.5
β
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
||γ||


Scatter of joint posterior on β, ||γ|| with prior
β
||γ||
30
25
20
15
10
5
0
−5
−10
−15
0
0.2
0.4
0.6
0.8
1
1.2
1.4
1.6
1.8
2

methods

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. The Figures illustrate how the use of the prior improves the convergence properties of the Gibbs sampler, by eliminating the extremely large draws of β.

*

References

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