Econometric Reviews
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/lecr20

Normalization in Econometrics

James D. Hamilton [a b], Daniel F. Waggoner [a b] & Tao Zha [a b]
[a] University of California, San Diego, California, USA
[b] Federal Reserve Bank of Atlanta, Atlanta, Georgia, USA
Published online: 04 May 2007.

To cite this article: James D. Hamilton , Daniel F. Waggoner & Tao Zha (2007) Normalization in Econometrics, Econometric Reviews, 26:2-4, 221-252, DOI: 10.1080/07474930701220329

To link to this article: http://dx.doi.org/10.1080/07474930701220329

Econometric Reviews, 26(2–4):221–252, 2007

ISSN: 0747-4938 print/1532-4168 online
DOI: 10.1080/07474930701220329

Taylor & Francis
Taylor & Francis Group

NORMALIZATION IN ECONOMETRICS

James D. Hamilton, Daniel F. Waggoner, and Tao Zha □ *University of California, San Diego, California, USA and Federal Reserve Bank of Atlanta, Atlanta, Georgia, USA*

□ *The issue of normalization arises whenever two different values for a vector of unknown parameters imply the identical economic model. A normalization implies not just a rule for selecting which among equivalent points to call the maximum likelihood estimate (MLE), but also governs the topography of the set of points that go into a small-sample confidence interval associated with that MLE. A poor normalization can lead to multimodal distributions, disjoint confidence intervals, and very misleading characterizations of the true statistical uncertainty. This paper introduces an identification principle as a framework upon which a normalization should be imposed, according to which the boundaries of the allowable parameter space should correspond to loci along which the model is locally unidentified. We illustrate these issues with examples taken from mixture models, structural vector autoregressions, and cointegration models.*

Keywords Cointegration; Local identification; Mixture distributions; Maximum likelihood estimate; Numerical Bayesian methods; Regime-switching; Small sample distributions; Vector autoregressions; Weak identification.

JEL Classification C1; C32.

1. INTRODUCTION

An econometric model is said to be unidentified if two or more values for a vector of unknown parameters imply the identical probability law. In traditional discussions of the identification problem (Koopmans, 1953, for example), these different parameters could have very different economic implications. A good deal of effort is accordingly devoted to ensuring the validity of the identifying restrictions that led to selecting one parameter value over the others.

On the other hand, when two or more parameter values are observationally equivalent and furthermore have the identical economic implications, we say that the model simply requires a normalization. One's

Received September 15, 2005; Accepted June 13, 2006
Address correspondence to Tao Zha, Federal Reserve Bank of Atlanta, 1000 Peachtree Street NE, Atlanta, GA 30309, USA; E-mail: tzha@earthlink.net

first thought might be that if "normalization" is defined as an identification restriction that has no implications for statements we would make about economic behavior, it should not make any difference how one deals with it. However, a host of counterexamples discussed in this paper show that in fact it can matter a great deal.

The fact that normalization can materially affect the conclusions one draws with likelihood-based methods is not widely recognized. Here the normalization problem arises when the likelihood $f(\mathbf{y};\boldsymbol{\theta}_1) = f(\mathbf{y};\boldsymbol{\theta}_2)$ for all possible values of $\mathbf{y}$. Since $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ imply the identical observed behavior and since the maximum likelihood estimates (MLEs) themselves are invariant with respect to a reparameterization, many scholars (e.g., Anderson, 2005) have assumed that the question of normalization was irrelevant as long as parameters are estimated by the full information maximum likelihood (FIML) or limited information maximum likelihood (LIML) method. If one were interested in reporting only the MLE and the probability law that it implies for $\mathbf{y}$, this would indeed be the case. The problem arises when one wishes to go further and make a statement about a region of the parameter space around $\boldsymbol{\theta}_1$, for example, in constructing confidence sets. In this case, normalization is not just a rule for selecting $\boldsymbol{\theta}_1$ over $\boldsymbol{\theta}_2$ but in fact becomes a rule for selecting a whole region of points $\{\boldsymbol{\theta}_1^* : \boldsymbol{\theta}_1^* \in \Omega(\boldsymbol{\theta}_1)\}$ to associate with $\boldsymbol{\theta}_1$. Our paper shows that a poor normalization can have the consequence that two nearly observationally equivalent probability laws ($f(\mathbf{y};\boldsymbol{\theta}_1)$ arbitrarily close to $f(\mathbf{y};\boldsymbol{\theta}_1^*)$) are associated with widely different points in the parameter space ($\boldsymbol{\theta}_1$ arbitrarily far from $\boldsymbol{\theta}_1^*$). The result can be a multimodal distribution for the MLE $\hat{\boldsymbol{\theta}}$ that is grossly misrepresented by a simple mean and variance. More fundamentally, the economic interpretation one places on the region $\Omega(\boldsymbol{\theta}_1)$ is inherently problematic in such a case.

This problem has previously been recognized in a variety of individual settings, but to our knowledge with no unifying treatment of the general nature of the problem and its solution. The contribution of the present paper is to articulate the general statistical principles underlying the problems heretofore raised in isolated literatures. We propose a general solution to the normalization problem, following an "identification principle." We contrast our work with previous research dealing with identification problems.

The conventional way of implementing normalization is by means of a priori restrictions to restrict the parameter space in certain ways. One could also implement normalization as part of the Bayesian prior, as we will show in Section 3. Such a prior is likely to be nonstandard and may have material consequences on the marginal likelihood that complicate the model comparison. Our purpose is to show that contrary to the common belief, normalization affects inferential conclusions

for quantifying the uncertainty about the MLEs, no matter whether normalization is implemented by a priori restrictions or Bayesian prior probability density functions (pdfs).

We find it helpful to introduce the key issues in Section 2 with a simple example, namely estimating the parameter σ for an i.i.d. sample of $N(0, \sigma^2)$ variables. Section 3 illustrates how the general principles proposed in Section 2 apply in mixture models. Section 4 discusses structural vector autoregressions (VARs), while Section 5 investigates cointegration. A number of other important econometric models could also be used to illustrate these principles, but are not discussed in detail in this paper. These include binary response models, where one needs to normalize coefficients in a latent process or in expressions that only appear as ratios (e.g., Hauck Jr. and Donner, 1977; Manski, 1988), dynamic factor models, where the question is whether a given feature of the data is mapped into a parameter of factor i or factor j (e.g., Otrok and Whiteman, 1998); and neural networks, where the possibility arises of hidden unit weight interchanges and sign flips (e.g., Chen et al., 1993; Rüger and Ossen, 1996). Section 6 summarizes our practical recommendations for applied research in any setting requiring normalization.

2. NORMALIZATION AND AN IDENTIFICATION PRINCIPLE

2.1. Motivation

We can illustrate the key issues associated with normalization through the following example. Suppose $y_t = \sigma\varepsilon_t$ where $\varepsilon_t \sim$ i.i.d. $N(0,1)$. Denote the likelihood function with the sample size T by $f(\mathbf{y};\sigma)$, where $\mathbf{y} = \{y_1, \ldots, y_T\}$. It follows that the log likelihood function is

$$\log f(\mathbf{y};\sigma) = -(T/2)\log(2\pi) - (T/2)\log(\sigma^2) - \sum_{t=1}^{T} y_t^2/(2\sigma^2).$$

The likelihood function is of course a symmetric function of σ, with positive and negative values of σ implying identical probabilities for observed values of $\mathbf{y}$. One needs to restrict σ further than just $\sigma \in \Re^1$ in order to infer the value of σ from observation of $\mathbf{y}$. The obvious (and, we will argue, correct) normalization is to impose the restriction $\sigma > 0$. But consider the consequences of using some alternative rule for normalization, such as $\sigma \in A = \{(-2, 0) \cup [2, \infty)\}$. This also would technically solve the normalization problem, in that distinct elements of A imply different probability laws for y_t. But inference about σ that relies on this normalization runs into three potential pitfalls.

First, the Bayesian posterior distribution $\pi(\sigma \mid \mathbf{y})$ is bimodal and classical confidence regions are disjoint. This might not be a problem as long as

one accurately reported the complete distribution. However, if we had generated draws numerically from $\pi(\sigma \mid \mathbf{y})$ and simply summarized this distribution by its mean and standard deviation (as is often done in more complicated, multidimensional problems), we would have a grossly misleading inference about the nature of the information contained in the sample about σ.

Second, the economic interpretation one places on σ is fundamentally different over different regions of A, separated by the point $\sigma = 0$ at which the log likelihood is $-\infty$. In the positive region, higher values of σ imply more variability of y_t, whereas in the negative region, higher values of σ imply less variability of y_t. If one had adopted the $\sigma \in A$ normalization, the question of whether σ is large or small would not be of fundamental interest, and why a researcher would even want to calculate the posterior mean and standard deviation of σ is not at all clear.

Third, the economic interpretation one places on the interaction between variables is fundamentally different over different regions of A. In VAR analysis, a common goal is to estimate the effect of shocks on the variables in the system. For this example, the impulse response function is simply

$$\partial y_{t+j}/\partial \varepsilon_t = \begin{cases} \sigma & j = 0 \\ 0 & j = 1, 2, \ldots \end{cases}.$$

Thus the consequences of a one unit increase in ε_t are different over different regions of the parameter space. In the positive region, a positive shock to ε_t is interpreted as something that increases y_t, whereas over the negative region, a positive shock to ε_t is interpreted as something that decreases y_t. Again, if this is the normalization one had imposed, it is not clear why one would ever want to calculate an object such as $\partial y_{t+j}/\partial \varepsilon_t$.

In this example, these issues are sufficiently transparent that no researcher would ever choose such a poor normalization or fall into these pitfalls. However, we will show below that it is very easy to make similar kinds of mistakes in a variety of more complicated econometric contexts. Before doing so, we outline the general principles that we propose as a guideline for the normalization question in any setting.

2.2. General Principles

Our starting point is the observation that the normalization problem is fundamentally a question of identification. Let $\boldsymbol{\theta} \in \Re^k$ denote the parameter vector of interest and $f(\mathbf{y}; \boldsymbol{\theta})$ the likelihood function. Following Rothenberg (1971), two parameter points $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are said to be observationally equivalent if $f(\mathbf{y}; \boldsymbol{\theta}_1) = f(\mathbf{y}; \boldsymbol{\theta}_2)$ for all values of $\mathbf{y}$.

The structure is said to be globally identified at the point $\boldsymbol{\theta}_0$ if there is no other allowable value for $\boldsymbol{\theta}$ that is observationally equivalent to $\boldsymbol{\theta}_0$. The structure is said to be locally identified at $\boldsymbol{\theta}_0$ if there exists an open neighborhood around $\boldsymbol{\theta}_0$ containing no other value of $\boldsymbol{\theta}$ that is observationally equivalent to $\boldsymbol{\theta}_0$.

In the absence of a normalization condition, the structure would typically be globally unidentified but locally identified. The two points implying identical observed behavior ($\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) are typically separated in $\Re^k$. However, usually there will be loci in $\Re^k$ along which the structure is locally unidentified or the log likelihood diverges to $-\infty$. These loci characterize the boundaries across which the interpretation of parameters necessarily changes. In the example presented in Section 2.1, since $k = 1$, the locus is simply a point in $\Re^1$, namely, $\sigma = 0$. The two points $\sigma_1 = 3$ and $\sigma_2 = -3$, for instance, are separated in $\Re^1$ and have the same likelihood value but with different economic interpretations in terms of impulse responses.

The normalization problem is to restrict $\boldsymbol{\theta}$ to a subset A of $\Re^k$. Our proposal is that the boundaries of A should correspond to the loci along which the structure is locally unidentified or the log likelihood is $-\infty$. We describe this as choosing a normalization according to an *identification principle*. In the previous simple example, using this locus as the boundary for A means defining A by the condition $\sigma > 0$ – the common-sense normalization for this transparent case.

One easy way to check whether a proposed normalization set A conforms to this identification principle is to make sure that the model is locally identified at all interior points of A. If it is not, then the normalization does not satisfy the identification principle. The following sections illustrate these ideas in a number of different settings with more complicated examples.

3. MIXTURE MODELS

One class of models for which the normalization problem arises is when the observed data come from a mixture of different distributions or regimes, as in the Markov-switching models proposed by Hamilton (1989). Consider for illustration the simplest i.i.d. mixture model, in which y_t is drawn from a $N(\mu_1, 1)$ distribution with probability p and a $N(\mu_2, 1)$ distribution with probability $1 - p$, so that its density is

$$f(y_t; \mu_1, \mu_2, p) = \frac{p}{\sqrt{2\pi}} \exp\left[\frac{-(y_t - \mu_1)^2}{2}\right] + \frac{1-p}{\sqrt{2\pi}} \exp\left[\frac{-(y_t - \mu_2)^2}{2}\right]. \quad (1)$$

The model is unidentified in the sense that, if one switches the labels for regime 1 and regime 2, the value of the likelihood function is

unchanged: $f(y_t; \mu_1, \mu_2, p) = f(y_t; \mu_2, \mu_1, 1-p)$. Before we can make any inference about the value of $\boldsymbol{\theta} = (\mu_1, \mu_2, p)'$ we need a resolution of this "label-switching" problem. Treatments of this problem include Celeux et al. (2000), Stephens (2000), and Frühwirth-Schnatter (2001).

How we choose to resolve the problem depends in part on why we are interested in the parameters in the first place. One possibility is that (1) is simply proposed as a flexible representation of the density of y_t. Here one has no interest in the value of $\boldsymbol{\theta}$ itself, but only in the shape of the distribution $f(\cdot)$. If this is one's goal, the best approach may be to simulate the posterior distribution of $\boldsymbol{\theta}$ without imposing any normalization at all while making sure that the full range of permutations gets sampled, and checking to make sure that the inferred distribution is exactly multimodally symmetric (e.g., Celeux et al., 2000). This can be more difficult to implement than it sounds, particularly if one tries to apply it to higher-dimensional problems. However, once the unrestricted multimodal distribution is successfully obtained, as long as one is careful to use this distribution only for purposes of making calculations about $f(\cdot)$, the multimodality of the distribution and ambiguity about the nature of $\boldsymbol{\theta}$ need not introduce any problems.

A second reason one might be interested in this model is as a structural description of a particular economic process for which the parameters $\boldsymbol{\theta}$ have clear and distinct economic interpretations. For example, y_t might be the value of GDP growth in year t, μ_1 the growth rate in expansions, μ_2 the growth rate in recessions, and p the probability of an expansion. In this case, the structural interpretation dictates the normalization rule that should be adopted, namely $\mu_1 > \mu_2$. A nice illustration and extension of this idea is provided by Smith and Summers (2003).

A third case is where the researcher believes that there is an underlying structural mechanism behind the mixture distribution, but its nature is not currently understood. For example, y_t might be an interest rate. The two means might be revealed in later research to be related to economic expansions and contractions, or to changes in policy, but the nature of regimes is not known a priori. For this case, the researcher believes that there exists a unique true value of $\boldsymbol{\theta}_0$. The goal is to describe the nature of the two regimes, e.g., one regime is characterized by 4% higher interest rates on average, for which purposes point estimates and standard errors for $\boldsymbol{\theta}$ are desired. One needs to restrict the space of allowed values of $\boldsymbol{\theta}$ to an identified subspace in order to be able to do that.

One way one might choose to restrict the space would be to specify $p > .5$, as in Aitkin and Rubin (1985) or Lenk and DeSarbo (2000). However, according to the identification principle discussed in Section 2, this is not a satisfactory solution to the normalization problem. This is because even if one restricts $p > .5$, the structure is still locally unidentified at any point

at which $\mu_1 = \mu_2$, for at any such point the likelihood function does not depend on the value of p.

To illustrate what difference the choice of normalization makes for this example, we calculated the log likelihood for a sample of 50 observations from the above distribution with $\mu_1 = 1, \mu_2 = -1$, and $p = .8$. Figure 1 plots contours of the log likelihood as a function of μ_1 and μ_2 for alternative values of p. The maximum value for the log likelihood (-79) is achieved near the true values, as shown in the upper left panel. The lower right panel is its exact mirror image, with a second maximum occurring near $\mu_1 = -1$, $\mu_2 = 1$, and $p = .2$. In the middle right panel ($p = .5$), points above the 45° line are the mirror image of those below. The proposed normalization ($p > .5$) restricts the space to the first

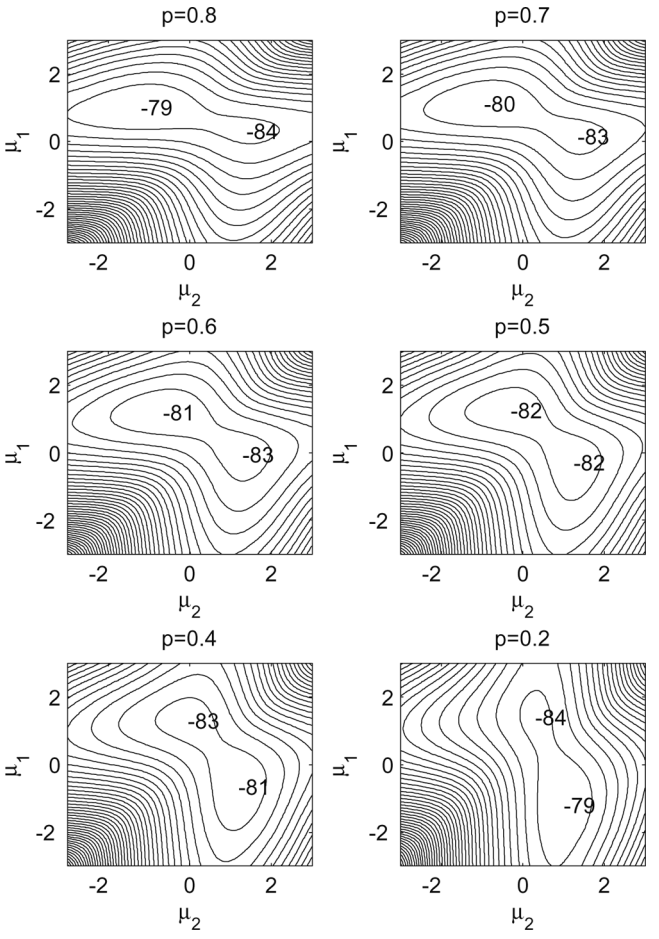

FIGURE 1 Contours of log likelihood for an i.i.d. Gaussian mixture of $T = 50$ observations. True values: $\mu_1 = 1, \mu_2 = -1, p = .8$.

three panels. This solves the normalization problem in the sense that there is now a unique global maximum to the likelihood function, and any distinct values of $\boldsymbol{\theta}$ within the allowable space imply different probability laws for y_t. However, by continuity of the likelihood surface, each of these panels has a near symmetry across the 45° line that is an echo of the exact symmetry of the $p = .5$ panel. Conditional on any value of p, the normalization $p > .5$ therefore results in one mass of probability centered at $\mu_1 = 1$, $\mu_2 = -1$, and a second smaller mass centered at $\mu_1 = -1$, $\mu_2 = 1$. Hence, although restricting $p > .5$ can technically solve the normalization problem, it does so in an unsatisfactory way. The problem arises because points interior to the normalized region include the axis $\mu_1 = \mu_2$, along which the labelling of regimes could not be theoretically defined, and across which the substantive meaning of the regimes switches.[1]

An alternative normalization would set $\mu_1 > \mu_2$, defining the allowable parameter space by the upper left triangle of all panels, choosing the locus along which the model is locally unidentified ($\mu_1 = \mu_2$) as the boundary for the parameter space. Note that over this region, the global likelihood surface is much better behaved.

To investigate this in further detail, we calculated the Bayesian posterior distributions. For a Bayesian prior we specified $\mu_i \sim N(0, 5)$ (with μ_1 independent of μ_2) and used a uniform prior for p. We will comment further on the role of these priors below. Appendix A describes the specifics of the Gibbs sampler used to simulate draws from the posterior distribution of $\boldsymbol{\theta}$. For each draw of $\boldsymbol{\theta}^{(i)}$, we kept $\boldsymbol{\theta}^{(i)}$ if $p^{(i)} > .5$, but used $(\mu_2^{(i)}, \mu_1^{(i)}, 1 - p^{(i)})'$ otherwise. We ran the Gibbs sampler for 5,500 iterations on each sample, with parameter values initialized from the prior, discarded the first 500 iterations, and interpreted the last 5,000 iterations as draws from the posterior distribution of parameters for that sample.[2] We repeated this process on 1,000 different samples of size $T = 10, 20, 50$, and 100. For the nth generated sample, we calculated the difference between the posterior mean $E(\boldsymbol{\theta} \mid \mathbf{y}^{(n)})$ and true value $\boldsymbol{\theta} = (1, -1, 0.8)'$. The mean squared errors across samples n are plotted as a function of the sample size T in the first row of Figure 2. The $\mu_1 > \mu_2$ normalization produces lower mean squared errors for any sample size for either of the mean parameters, substantially so for μ_2.

[1]This observation that simply restricting $\boldsymbol{\theta}$ to an identified subspace is not a satisfactory solution to the label-switching problem has also been forcefully made by Celeux et al. (2000), Stephens (2000), and Frühwirth-Schnatter (2001), though none of them interpret this problem in terms of the identification principle articulated here. Frühwirth-Schnatter suggested plotting the posterior distributions under alternative normalizations to try to find one that best respects the geometry of the posterior. Celeux et al. (2000) and Stephens (2000) proposed a decision-theoretic framework.

[2]Given the nature of i.i.d. draws, this number of iterations is more than sufficient for obtaining the accurate posterior mean.

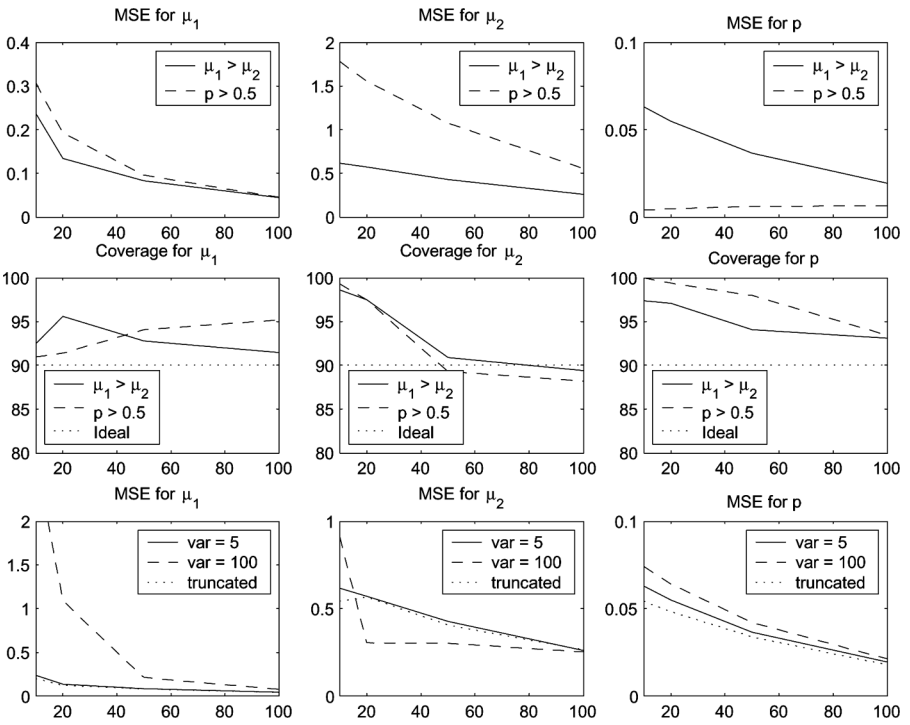

FIGURE 2 Performance of estimators with sample size on the horizontal axis. Top row: average squared difference between posterior mean and true value for indicated parameter when normalized by $\mu_1 > \mu_2$ (solid line) or by $p > .5$ (dashed line). First column: performance for estimate of μ_1; second column: performance for estimate of μ_2; third column: performance for estimate of p. Second row: ninety-percent coverage probabilities for indicated parameter when normalized by $\mu_1 > \mu_2$ (solid line) or by $p > .5$ (dashed line); dotted line gives nominal 90% goal. Third row: average squared difference between posterior mean and true value for indicated parameter when normalized by $\mu_1 > \mu_2$ for three different priors. Solid line: prior variance = 5; dashed line: prior vairance = 100; dotted line: truncated Normal prior.

Another key question is whether the posterior distributions accurately summarize the degree of objective uncertainty about the parameters. For each sample, we calculated a 90% confidence region for each parameter as implied by the Bayesian posterior distribution. We then checked whether the true parameter value indeed fell within this region, and calculated the fraction of samples for which this condition was satisfied. The second row in Figure 2 reports these 90% coverage probabilities for the two normalizations. The $\mu_1 > \mu_2$ normalization produces the most accurately sized test of the hypotheses $\mu_1 = \mu_{10}$ or $\mu_2 = \mu_{20}$ for samples with $T \geq 50$.

The $p > .5$ normalization does achieve a significantly better MSE for purposes of estimating p (upper right panel of Figure 2). However, this appears to be primarily a consequence of the prior forcing the estimate to be in the vicinity of its true value. The MSE for p for the $p > .5$

normalization actually deteriorates as the sample size T increases, and the coverage probabilities are quite poor (second row, third column of Figure 2).

On the other hand, the normalization $\mu_1 > \mu_2$ can also interact with the original symmetric prior for μ to substantially improve the accuracy of the prior information.[3]

If the original symmetric prior is

$$\begin{bmatrix}\mu_1\\ \mu_2\end{bmatrix} \sim N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}\varsigma^2 & 0\\ 0 & \varsigma^2\end{bmatrix}\right), \tag{2}$$

then $E(\mu_1^* = \max\{\mu_1, \mu_2\}) = \varsigma/\sqrt{\pi}$.[4] For the prior used in the above calculations, $\varsigma = \sqrt{5}$. Hence the prior expectation of μ_1^* is 1.26, and likewise $E(\mu_2^*) = -1.26$, both close to the true values of ± 1. To see how the prior can adversely interact with normalization, suppose instead we had set $\varsigma^2 = 100$. In the absence of normalization, this would be an attractive uninformative prior. With the normalization $\mu_1 > \mu_2$, however, it implies a prior expectation $E(\mu_1^*) = 5.64$ and a nearly even chance that μ_1^* would exceed this value, even though in 100,000 observations on y_t, one would not be likely to observe a single value as large as this magnitude that is proposed as the *mean* of one of the subpopulations.[5] Likewise the prior is also assigning a 50% probability that $\mu_2 < -5$, when the event $y_t < -5$ is also virtually impossible.

The third row in Figure 2 compares mean squared errors that would result from the $\mu_1 > \mu_2$ normalization under different priors. Results for the $N(0, 5)$ prior are represented by the solid lines. This solid line in the left panel of the third row in Figure 2 is identical to the solid line in the left panel of the first row in Figure 2, but the scale is different in order to try to convey the huge mean squared errors for μ_1 that result under the $N(0, 100)$ prior (the latter represented by the dashed line in the third row of Figure 2). Under the $N(0, 100)$ prior, the $\mu_1 > \mu_2$ normalization does a substantially worse job at estimating μ_1 or p than would the $p > .5$ normalization for sample sizes below 50. Surprisingly, it does a better job at estimating μ_2 for moderate sample sizes precisely because the strong bias introduced by the prior offsets the bias of the original estimates.

It is clear from this discussion that we need to be aware not only of how the normalization conforms to the topography of the likelihood function, but also with how it interacts with any prior that we might use in Bayesian analysis. Given the normalization $\mu_1 > \mu_2$, rather than the prior (2), it

[3]By symmetric we mean the prior pdf is the same or symmetric for both μ_1 and μ_2.

[4]See Ruben (1954, Table 2).

[5]The probability that a variable drawn from the distribution with the larger mean ($N(1, 1)$) exceeds 5.5 is .00000340.

seems better to employ a truncated Gaussian prior, where $\mu_1 \sim N(\bar{\mu}_1, \varsigma_1^2)$ and

$$\pi(\mu_2 \mid \mu_1) = \begin{cases} \dfrac{1}{\Phi[(\mu_1 - \bar{\mu}_2)/\varsigma_2]\sqrt{2\pi}\varsigma_2} \exp\left(\dfrac{-(\mu_2 - \bar{\mu}_2)^2}{2\varsigma_2^2}\right) & \text{if } \mu_2 \leq \mu_1 \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

for $\Phi(z) = \text{Prob}(Z \leq z)$ for $Z \sim N(0,1)$. Here $\bar{\mu}_2$ and ς_2^2 denote the mean and variance of the distribution that is truncated by the condition $\mu_2 < \mu_1$. One drawback of this truncated Gaussian prior is that it is no longer a natural conjugate for the likelihood, and so the Gibbs sampler must be adapted to include a Metropolis–Hastings step rather than a simple draw from a normal distribution, as detailed in Appendix A.

We redid the above analysis using this truncated Gaussian prior with $\bar{\mu}_1 = \bar{\mu}_2 = 0$ and $\varsigma_1^2 = \varsigma_2^2 = 5$. When $\mu_1 = 0$, for example, this prior implies an expected value for μ_2 of $\bar{\mu}_2 + \varsigma_2 M_2 = -1.78$ where $M_2 = -\phi(c_2)/\Phi(c_2) = .7979$ with $c_2 = (\mu_1 - \bar{\mu}_2)/\varsigma_2 = 0$ and a variance for μ_2 of $\varsigma_2^2[1 - M_2(M_2 - c_2)] = 1.82$.[6] Mean squared errors resulting from this truncated Gaussian prior are reported in the dotted lines in the third row of Figure 2. These uniformly dominate those for the simple $N(0,5)$ prior.

To summarize, the $p > .5$ normalization introduces substantial distortions in the Bayesian posterior distribution that can be largely avoided with the $\mu_1 > \mu_2$ normalization. Normalization based on the identification principle seems to produce substantially superior point estimates of μ_1 and μ_2 for small samples and much better coverage probabilities for larger samples. The exercises performed in this section also show that normalization can be viewed as part of the prior pdf. An "unreasonable" prior or normalization can distort likelihood-based inferences about the point estimates. Although normalization can be in principle implemented via a Bayesian prior pdf, such a prior is likely to be nonstandard and thus its small-sample statistical properties can be sensitive to how it is designed. The following section presents this kind of situation.

4. STRUCTURAL VARS

Let $\mathbf{y}_t$ denote an $(n \times 1)$ vector of variables observed at date t. Consider a structural VAR of the form

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{k} + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \mathbf{u}_t \tag{4}$$

where $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{D}^2)$ with $\mathbf{D}$ a diagonal matrix. A structural VAR typically imposes both exclusion restrictions and normalization conditions on $\mathbf{B}_0$

[6]See for example Maddala (1983, pp. 365–366).

in order to be identified. To use a familiar example (Hamilton, 1994, pp. 330–331), let $\mathbf{y}_t = (q_t, p_t, w_t)'$ where q_t denotes the log of the number of oranges sold in year t, p_t the log of the price, and w_t the number of days with below-freezing temperatures in Florida (a key orange-producing state) in year t. We are interested in a demand equation of the form

$$q_t = \beta p_t + \boldsymbol{\delta}_1' \mathbf{x}_t + u_{1t} \tag{5}$$

where $\mathbf{x}_t = (1, \mathbf{y}_{t-1}', \mathbf{y}_{t-2}', \ldots, \mathbf{y}_{t-p}')'$ and the demand elasticity β is expected to be negative. Quantity and price are also determined by a supply equation,

$$q_t = \gamma p_t + h w_t + \boldsymbol{\delta}_2' \mathbf{x}_t + u_{2t},$$

with the supply elasticity expected to be positive ($\gamma > 0$) and freezing weather to discourage orange production ($h < 0$). We might also use an equation for weather of the form $w_t = \boldsymbol{\delta}_3' \mathbf{x}_t + u_{3t}$, where perhaps $\boldsymbol{\delta}_3 = \mathbf{0}$. This system is an example of (4) incorporating both exclusion restrictions (weather does not affect demand directly, and neither quantity nor price affect the weather) and normalization conditions (three of the elements of $\mathbf{B}_0$ have been fixed at unity):

$$\mathbf{B}_0 = \begin{bmatrix} 1 & -\beta & 0 \\ 1 & -\gamma & -h \\ 0 & 0 & 1 \end{bmatrix}. \tag{6}$$

The latter seems a sensible enough normalization, in that the remaining free parameters (β, γ, and h) are magnitudes of clear economic interpretation and interest. However, the identification principle suggests that it may present problems, in that the structure is unidentified at some interior points in the parameter space. Specifically, at $h = 0$, the value of the likelihood would be unchanged if β were switched with γ. Moreover, the log likelihood approaches $-\infty$ as $\beta \to \gamma$.

To see the practical consequences of this, consider the following parametric example:

$$\begin{bmatrix} 1 & 2 & 0 \\ 1 & -0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \begin{bmatrix} 0.8 & 1.6 & 0 \\ 1.2 & -0.6 & 0.6 \\ 0 & 0 & 1.8 \end{bmatrix} \begin{bmatrix} q_{t-1} \\ p_{t-1} \\ w_{t-1} \end{bmatrix} + \begin{bmatrix} 0 & 0 & 0 \\ -0.8 & 0.4 & -0.4 \\ 0 & 0 & -0.9 \end{bmatrix} \begin{bmatrix} q_{t-2} \\ p_{t-2} \\ w_{t-2} \end{bmatrix} + \begin{bmatrix} u_{dt} \\ u_{st} \\ u_{wt} \end{bmatrix}. \tag{7}$$

In this example, the true demand elasticity $\beta = -2$ and supply elasticity $\gamma = .5$, while $h = -.5$ and $\mathbf{D} = \mathbf{I}_3$. Demand shocks are AR(1) with

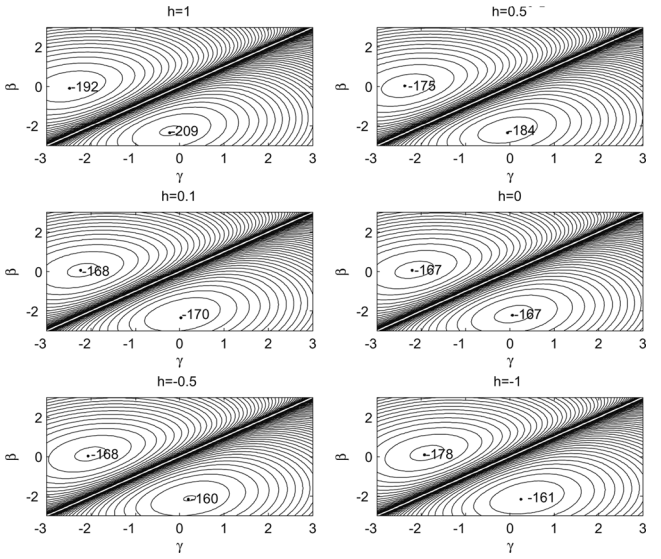

FIGURE 3 Contours of concentrated log likelihood of structural VAR under the β-normalization.

exponential decay factor .8 while supply and weather shocks are AR(2) with damped sinusoidal decay.

Figure 3 shows contours of the concentrated log likelihood for a sample of size $T = 50$ from this system.[7] Each panel displays contours of $\mathcal{L}(\beta, \gamma, h)$ as functions of β and γ for selected values of h. The middle right panel illustrates both problems with this normalization noted above: when $h = 0$, the likelihood function is unchanged when β is switched with γ. Furthermore, the log likelihood is $-\infty$ along the locus $\beta = \gamma$, which partitions this panel into the two regions that correspond to identical values for the likelihood surface.

The global maximum for the likelihood function occurs at $\beta = -2.09$, $\gamma = .28$, and $h = -.69$, corresponding to the hill in the lower right triangle of the bottom left panel in Figure 3. Although the upper left triangle is not the mirror image of the lower right in this panel, it nevertheless is the case that, even at the true value of h, the likelihood function is characterized by two separate concentrations of mass, one around the true values $(\beta = -2, \gamma = .5)$ and a second smaller mass around their flipped values $(\beta = .5, \gamma = -2)$. Although the likelihood values associated with the former are much larger than the latter, the likelihood function merges

[7]The likelihood has been concentrated by first regressing q_t and p_t on $\mathbf{y}_{t-1}$ and $\mathbf{y}_{t-2}$, and regressing w_t on w_{t-1} and w_{t-2}, to get a residual vector $\hat{\mathbf{u}}_t$ and then evaluating at the true $\mathbf{D} = \mathbf{I}_3$. That is, for $\mathbf{B}_0(\beta, \gamma, h)$ the matrix in (6), we evaluated

$$\mathcal{L}(\beta, \gamma, h) = -1.5T \ln(2\pi) + (T/2) \ln(|\mathbf{B}_0|^2) - (1/2) \sum_{t=1}^{T} (\mathbf{B}_0 \hat{\mathbf{u}}_t)' (\mathbf{B}_0 \hat{\mathbf{u}}_t).$$

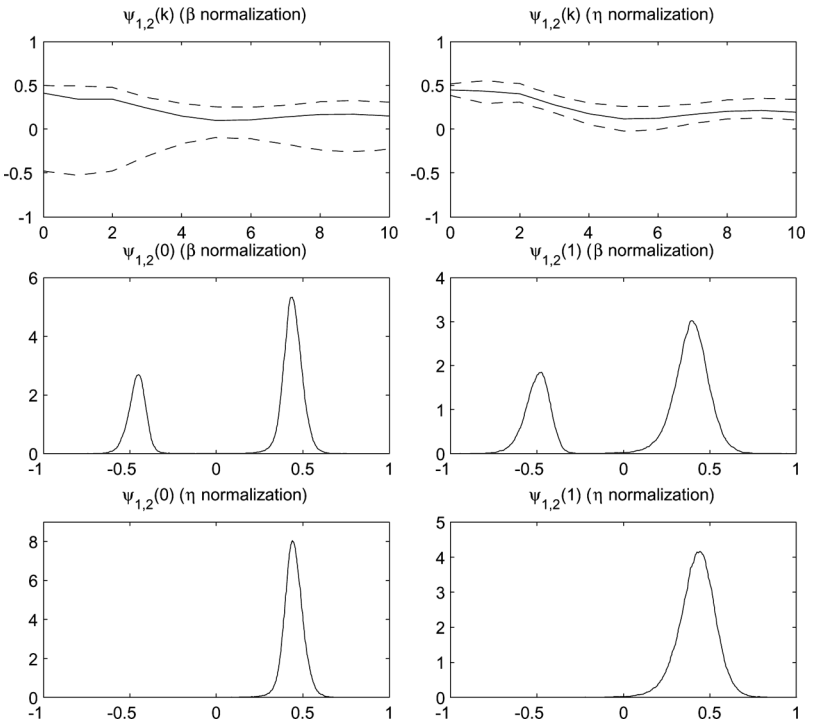

FIGURE 4 Top row: impulse-response function and 90% confidence interval for the effect of a one standard deviation increase in quantity demanded on the price k periods later under the β-normalization (left panel) and η-normalization (second panel). Second row: posterior density for the $k = 0$ (left panel) and $k = 1$ (right panel) values of the impulse-response function plotted in the upper left panel. Third row: posterior density for the $k = 0$ and $k = 1$ values of the impulse-response function plotted in the upper right panel.

continuously into the exact mirror image case as h approaches zero, at which the masses become identical. Because the likelihood function is relatively flat with respect to h, the result is a rather wild posterior distribution for impulse responses under this normalization.

To describe this distribution systematically, we generated 100,000 draws from the posterior distribution of $(\beta, \gamma, h, d_1, d_2, d_3 \,|\, \mathbf{y}_1, \ldots, \mathbf{y}_T)$ for a representative sample with $T = 50$ and with a flat prior.[8] The 95% confidence interval for β over these 100,000 draws is the range $[-11.3, +5.5]$. A particularly wild impulse response function $\psi_{ij}(k) = \partial y_{j,t+k}/\partial u_{it}$ is that for $\psi_{12}(k)$, the effect of a demand shock on price. The mean value and 90% confidence intervals are plotted as a function of k in the left panel of the first row in Figure 4. It is instructive (though

[8]See Appendix B for details on the algorithm used to generate these draws. For this kind of model, this number of Markov Chain Monte Carlo draws is sufficient to guarantee convergence as shown in Waggoner and Zha (2003a).

not standard practice) to examine the actual probability distribution underlying this familiar plot. The left panel of the second row in Figure 4 shows the density of $\psi_{12}(0)$ across these 100,000 draws, which is curiously bimodal. That is, in most of the draws, a one standard deviation shock to demand is interpreted as something that raises the price by .5, though in a significant minority of the draws, a positive shock to demand is interpreted as something that lowers the price by .5. This ambiguity about the fundamental question being asked (what one means by a one-unit shock to demand) interacts with uncertainty about the other parameters to generate the huge tails for the estimated value of $\psi_{12}(1)$ (the right panel of the second row in Figure 4). We would opine that, even though the researcher's maximum likelihood estimates correctly characterize the true data-generating process, such empirical results could prove impossible to publish.

The identification principle suggests that the way to get around the problems revealed in Figure 3 is to take the $\beta = \gamma$ axis as a boundary for the normalized parameter space rather than have it cut through the middle. More generally, we seek a normalization for which the matrix $\mathbf{B}_0$ in (6) becomes non-invertible only at the boundaries of the region. Let $\mathbf{C}$ denote the first two rows and columns of $\mathbf{B}_0$:

$$\mathbf{C} = \begin{bmatrix} 1 & -\beta \\ 1 & -\gamma \end{bmatrix}.$$

We thus seek a normalization for which $\mathbf{C}$ is singular only at the boundaries. One can see what such a region looks like by assuming that $\mathbf{C}^{-1}$ exists and premultiplying (4) by

$$\begin{bmatrix} \mathbf{C}^{-1} & \mathbf{0} \\ \mathbf{0}' & 1 \end{bmatrix}.$$

We then have

$$\begin{bmatrix} 1 & 0 & \pi_1 \\ 0 & 1 & \pi_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \mathbf{\Pi}_1 \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \mathbf{y}_{t-2} + \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{bmatrix}. \tag{8}$$

Figure 5 plots likelihood contours for this parameterization as a function of π_1, π_2, and ρ, the correlation between v_{1t} and v_{2t}.[9] Although this is exactly the same sample of data displayed in Figure 3, the likelihood function for this parameterization is perfectly well behaved, with a unique mode near the population values of $\pi_1 = .4$, $\pi_2 = -.2$, and $\rho = -.51$.

[9]For this graph, we set $E(v_{1t}^2) = .68$ and $E(v_{2t}^2) = .32$, their population values.

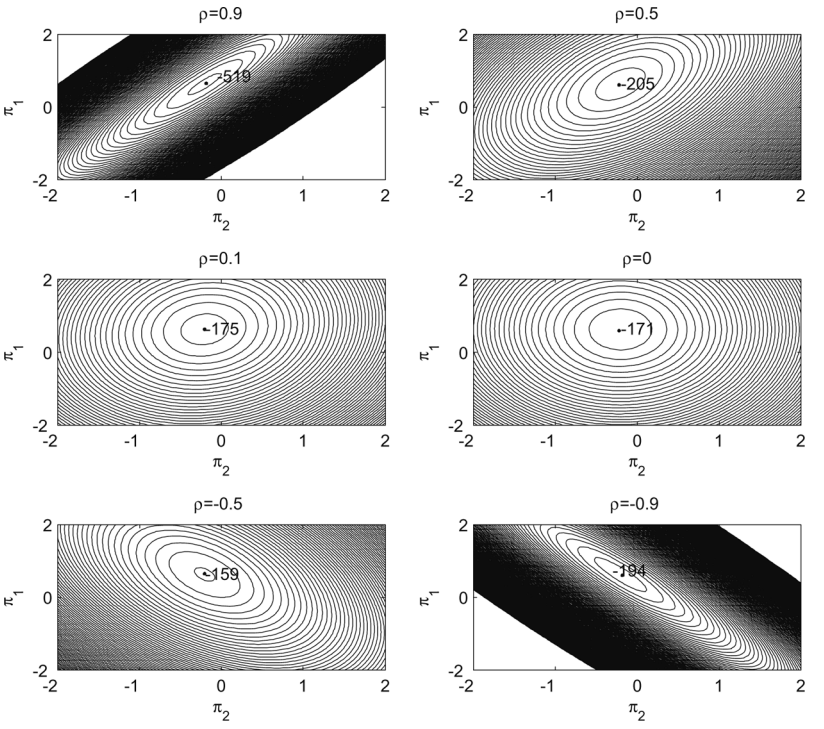

FIGURE 5 Contours of concentrated log likelihood of structural VAR under the π-normalization.

Indeed, (8) will be recognized as the reduced-form representation for this structural model as in Hamilton (1994, p. 245). The parameters all have clear interpretations and definitions in terms of basic observable properties of the data. The value of π_1 tells us whether the conditional expectation of q_t goes up or down in response to more freezing weather, π_2 does the same for p_t, and ρ tells us whether the residuals from these two regressions are positively or negatively correlated. Ninety-five percent confidence intervals from the same 100,000 draws described above are [.00, .71] for π_1 and [−.42, .04] for π_2.

Although this π-normalization eliminates the egregious problems associated with the β-normalization in (6), it cannot be used to answer all the original questions of interest, such as finding the value of the demand elasticity or the effects of a demand shock on price. We can nevertheless use the π-normalization to get a little more insight into why we ran into problems with the β-normalization. One can go from the π-normalization back to the β-normalization by premultiplying (8) by

$$\begin{bmatrix} \mathbf{C} & \mathbf{0} \\ \mathbf{0}' & 1 \end{bmatrix}$$

to obtain

$$\begin{bmatrix} 1 & -\beta & \pi_1 - \beta\pi_2 \\ 1 & -\gamma & \pi_1 - \gamma\pi_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \begin{bmatrix} v_{1t} - \beta v_{2t} \\ v_{1t} - \gamma v_{2t} \\ v_{3t} \end{bmatrix}. \tag{9}$$

Comparing (9) with (7), the structural parameter β must be chosen so as to make the (1, 3) element of $\mathbf{B}_0$ zero, or

$$\beta = \pi_1/\pi_2. \tag{10}$$

Given β, the parameter γ must be chosen so as to ensure $E(v_{1t} - \beta v_{2t})(v_{1t} - \gamma v_{2t}) = 0$, or

$$\gamma = \frac{\sigma_{11} - \beta\sigma_{12}}{\sigma_{12} - \beta\sigma_{22}}$$

for $\sigma_{ij} = E(v_{it}v_{jt})$. The value of h is then obtained from the (2, 3) element of $\mathbf{B}_0$ as

$$h = -(\pi_1 - \gamma\pi_2).$$

The problems with the posterior distribution for β can now be seen directly from (10). The data allow a substantial possibility that π_2 is zero or even positive, that is, that more freezes actually result in a lower price of oranges. Assuming that more freezes mean a lower quantity produced, if a freeze produces little change in price, the demand curve must be quite steep, and if the price actually drops, the demand curve must be upward sloping. A steep demand curve thus implies either a large positive or a large negative value for β, and when $\pi_2 = 0$, we switch from calling β an infinite positive number to calling it an infinite negative number. Clearly, a point estimate and standard error for β are a poor way to describe this inference about the demand curve. If π_2 is in the neighborhood of zero, it would be better to convey the apparent steepness of the demand curve by reparameterizing (6) as

$$\mathbf{B}_0 = \begin{bmatrix} -\eta & 1 & 0 \\ 1 & -\gamma & -h \\ 0 & 0 & 1 \end{bmatrix} \tag{11}$$

and concluding that η may be near zero.

When we performed the analogous 100,000 draws for the η-normalization (11), the 95% confidence interval for η is $[-1.88, .45]$, a more convenient and accurate way to summarize the basic fact that the demand curve is relatively steep, with elasticity $\beta = \eta^{-1} > -.53$ and possibly

even vertical or positively sloped. The response of price to a demand shock for this normalization is plotted in the right panel of the first row of Figure 4. The bimodality of the distribution of $\psi_{12}(0)$ and enormous tails of $\psi_{12}(1)$ have both disappeared (third row of Figure 4).

That such a dramatic improvement is possible from a simple renormalization may seem surprising, since for any given value for the parameter vector $\boldsymbol{\theta}$, the impulse-response function $\partial y_{j,t+k}/\partial u^*_{1t}$ for the η-normalization is simply the constant β^{-1} times the impulse-response function $\partial y_{j,t+k}/\partial u_{1t}$ for the β-normalization. Indeed, we have utilized this fact in preparing the right panel of the first row of Figure 4, multiplying each value of $\partial y_{2,t+k}/\partial u^*_{1t}$ by the constant $-.5$ before plotting the figure so as to get a value that corresponds to the identical concept and scale as the one measured in the left panel of the first row in Figure 4. The difference between this harmless rescaling (multiplying by the constant $-.5$) and the issue of normalization discussed in this paper is that the left panel of the first row of Figure 4 is the result of multiplying $\partial y_{2,t+k}/\partial u^*_{1t}$ not by the constant $-.5$ but rather by β^{-1}, which is a *different* magnitude for each of the 100,000 draws. Even though $\partial y_{2,t+k}/\partial u^*_{1t}$ is reasonably well-behaved across these draws, its product with β^{-1} is, as we see in the left panel of the first row of Figure 4, all over the map.

It is instructive also to examine the likelihood contours that result from the η-normalization which are plotted in Figure 6. The η-normalization does not satisfy the identification principle, because interior to the parameter space are the loci $\eta = \gamma^{-1}$ along which the model is locally unidentified. However, the height of the local hills that are across the $\eta\gamma = 1$ chasms relative to the global MLE is sufficiently below the height at the MLE (located near the center of the lower left panel) that the contribution of these problematic regions to the posterior distribution is negligible. Although the η-normalization does not literally satisfy the identification principle, for practical purposes it is sufficiently close (in the sense of relevant probability masses or likelihood values) to a true identification-based normalization that the problems associated with the β-normalization have been essentially eliminated.

This example also helps clarify the connection between the normalization issues raised here and those studied in previous treatments of the local identification problem. The local identification problem with $\beta = \gamma$ is in many respects similar to that in traditional simultaneous equations analysis, where the discussion of local non-identification can be found in Pagan and Robertson (1997), Staiger and Stock (1997), Hahn and Hausman (2002), Stock et al. (2002), Forchini and Hillier (2003), Yogo (2004), and Stock and Yogo (2005). Chao and Swanson (2005) adopted a classical point of view and Drèze (1976), Drèze and Morales (1976), Drèze and Richard (1983), Kleibergen and van Dijk (1998), and Chao and Phillips (2005) employed a Bayesian perspective.

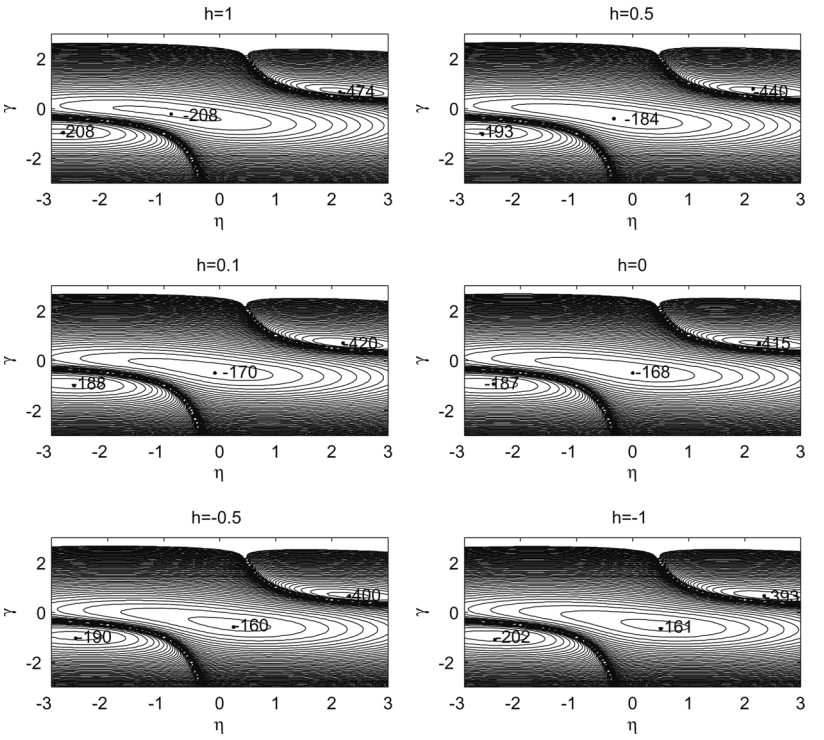

FIGURE 6 Contours of concentrated log likelihood of structural VAR under the η-normalizations.

The local identification problem in the classical analysis often relates to weak instruments. To see how normalization is related to this problem, suppose our goal was to estimate the price-sensitivity of demand using the two-stage least squares (2SLS) method, and considered the choice between estimating either (5) or the reverse regression,

$$p_t = \eta q_t + \boldsymbol{\psi}_1' \mathbf{x}_t + \varepsilon_{1t} \tag{12}$$

using $(w_t, \mathbf{x}_t')'$ as instruments in either case. Specification (12) is a preferred choice here because w_t serves as a better instrument for q_t in (12) than it does for p_t in (5). In other words, weak identification is less of a problem for 2SLS estimation of (12) than for (5). Exactly the same feature is relevant for our analysis, namely, weak identification turns out to be a less important feature in terms of probability mass of the likelihood surface as represented in Figure 6 than it is for the parameterization in Figure 3. However, the issue that results from this weak identification to which we're calling attention is different from the one with which the 2SLS estimator was concerned. The conventional focus was on how normalization can influence the parameter estimate itself. In our case,

because we are using FIML rather than 2SLS, the MLE's from the β-normalization are numerically identical to those for the η-normalization (that is, $\hat{\beta}$ is exactly equal to $\hat{\eta}^{-1}$). Even though weak identification does not affect the estimate in this case, we have just seen that it does affect the topology of the set of points that get associated with the respective MLE's for purposes of forming a confidence set.

The Bayesian analysis of simultaneous equations models is discussed by Drèze (1976), Drèze and Morales (1976), and Drèze and Richard (1983), among others, and is shown to lead to ill-behaved posterior distributions when a flat prior is used. This point is forcefully made by Kleibergen and van Dijk (1998), who suggest a prior directly on the reduced form of a simultaneous equations model with reduced-rank non-linear restrictions on the parameters.[10] They show how well the posterior distributions behave in this alternative framework. Similarly, for structural VARs, Sims and Zha (1994) show that a flat prior under the β-normalization leads to an improper posterior while a flat prior under the π-normalization leads to a well-behaved posterior distribution.[11]

Although the structural VARs studied here and the traditional simultaneous equations models have some common features, they are nonetheless different (Leeper et al., 1996). The most important difference is that the VAR analysis focuses on impulse responses to an economically interpretable shock. Even if the VAR model is *well identified* and one uses the standard informative prior (e.g., Litterman, 1986; Sims and Zha, 1998) to work directly on the reduced form, normalization is still needed (Sims and Zha, 1999). Sims and Zha (1999) and Waggoner and Zha (2003b) were the first to show that how the model is normalized has material consequences for the posterior distribution of impulse responses when the sample is small. As they showed, a solution to this problem cannot be resolved by existing methods proposed in the simultaneous equations literature.[12]

5. COINTEGRATION

Yet another instance where normalization can be important is in analysis of cointegrated systems. Consider

$$\Delta \mathbf{y}_t = \mathbf{k} + \mathbf{B}\mathbf{A}'\mathbf{y}_{t-1} + \boldsymbol{\zeta}_1 \Delta \mathbf{y}_{t-1} + \boldsymbol{\zeta}_2 \Delta \mathbf{y}_{t-2} + \cdots + \boldsymbol{\zeta}_{p-1} \Delta \mathbf{y}_{t-p+1} + \boldsymbol{\varepsilon}_t$$

where $\mathbf{y}_t$ is an $(n \times 1)$ vector of variables, $\mathbf{A}$ and $\mathbf{B}$ are $(n \times h)$ matrices of parameters of rank h, and $h < n$ is the number of cointegrating relations

[10] Chao and Phillips (2005) instead use a Jeffrey's prior.

[11] See Appendix B for more discussion.

[12] One could in principle design an informative prior directly on impulse responses to implement normalization, but there remain in practice economic and statistical issues that are yet to be sorted out (Sims, 2005).

among the variables in $\mathbf{y}_t$. Such models require normalization, since the likelihood function is unchanged if one replaces $\mathbf{B}$ by $\mathbf{BH}$ and $\mathbf{A}'$ by $\mathbf{H}^{-1}\mathbf{A}'$ for $\mathbf{H}$ any non-singular $(h \times h)$ matrix. Phillips (1994) studied the tendency (noted in a number of earlier studies cited in his article) for Johansen's (1988) normalization for the representation of a cointegrating vector to produce occasional extreme outliers, and explained how other normalizations avoid the problem by analyzing their exact small-sample distributions.

Two popular normalizations are to set the first h columns of $\mathbf{A}'$ equal to $\mathbf{I}_h$ (the identity matrix of dimension h) or to impose a length and orthogonality condition such as $\mathbf{A}'\mathbf{A} = \mathbf{I}_h$. However, both of these normalizations fail to satisfy the identification principle, because there exists an interior point in the allowable parameter space (namely, any point for which some column of $\mathbf{B}$ is the zero vector) at which the parameters of the corresponding row of $\mathbf{A}'$ become unidentified.

For illustration, consider a sample of $T = 50$ observations from the following model:

$$\begin{aligned} \Delta y_{1t} &= \varepsilon_{1t} \\ \Delta y_{2t} &= y_{1,t-1} - y_{2,t-1} + \varepsilon_{2t} \end{aligned} \tag{13}$$

with $\boldsymbol{\varepsilon}_t \sim N(\mathbf{0}, \mathbf{I}_2)$. This is an example of the above error-correction system in which $p = 1$, $\mathbf{B} = (0, b_2)', \mathbf{A}' = (a_1, a_2)$, and true values of the parameters are $b_2 = 1, a_1 = 1$, and $a_2 = -1$. The top panel of Figure 7 shows the consequences of normalizing $a_1 = 1$, displaying contours of the log likelihood as functions of a_2 and b_2. The global maximum occurs near the true values. However, as b_2 approaches zero, an iso-likelihood ellipse becomes infinitely wide in the a_2 dimension, reflecting the fact that a_2 becomes unidentified at this point. A similar problem arises along the a_1 dimension if one normalizes on $a_2 = 1$ (second panel). By contrast, the normalization $b_2 = 1$ does satisfy the identification principle for this example, and likelihood contours with respect to a_1 and a_2 (third panel) are well-behaved. This preferred normalization accurately conveys both the questions about which the likelihood is highly informative (namely, the fact that a_1 is the opposite value of a_2) and the questions about which the likelihood is less informative (namely, the particular values of a_1 or a_2).

For this numerical example, the identification is fairly strong in the sense that, from a classical perspective, the probability of encountering a sample for which the maximum likelihood estimate is in the neighborhood of $b_2 = 0$ is small, or from a Bayesian perspective, the posterior probability that b_2 is near zero is reasonably small. In such a case, the normalization $a_1 = 1$ or $a_2 = 1$ might not produce significant problems in practice. However, if the identification is weaker, the problems from a poor normalization can be

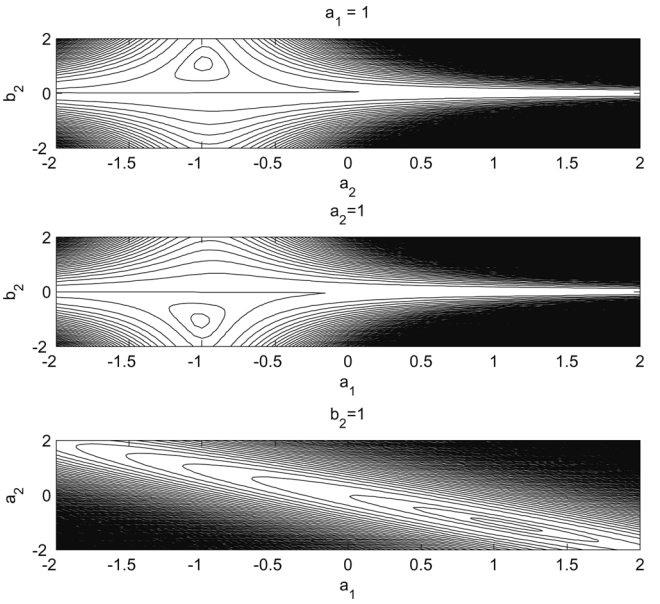

FIGURE 7 Contours of log likelihood for cointegrated system under three different normalizations.

much more severe. To illustrate this, we generated $N = 10,000$ samples each of size $T = 50$ from this model with $b_2 = .1$, $a_1 = 1$, and $a_2 = -1$, choosing the values of a_2 and b_2 for each sample so as to maximize the likelihood, given $a_1 = 1$. Row 1, column 1 panel of Figure 8 plots kernel estimates of the small-sample distribution of the maximum likelihood estimate $\hat{a}_2$. The distribution for $\hat{a}_2$ is extremely diffuse. Indeed, the MSE of $\hat{a}_2$ appears to be infinite, with the average value of $(\hat{a}_2 + 1)^2$ continuing to increase as we increased the number of Monte Carlo samples generated. The MSE is 208 when $N = 10,000$, with the smallest value generated being -665 and the biggest value 446. By contrast, if we normalize on $b_2 = 0.1$, the distributions of $\hat{a}_1$ and $\hat{a}_2$ are much better behaved (see the row 1, column 2 panel of Figure 8), with MSE's around .8.[13]

One can understand why the normalization that satisfies the identification principle ($b_2 = .1$) results in much better behaved estimates for this example by examining the reduced form of the model:

$$\begin{aligned} \Delta y_{1t} &= \varepsilon_{1t} \\ \Delta y_{2t} &= \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} + \varepsilon_{2t}. \end{aligned} \tag{14}$$

[13]Of course, normalizing $b_2 = 1$ (as one would presumably do in practice, not knowing the true b_2^0) would simply result in a scalar multiple of these distributions. We have normalized here on the true value ($b_2 = 0.1$) in order to keep the scales the same when comparing parameter estimates under alternative normalization schemes.

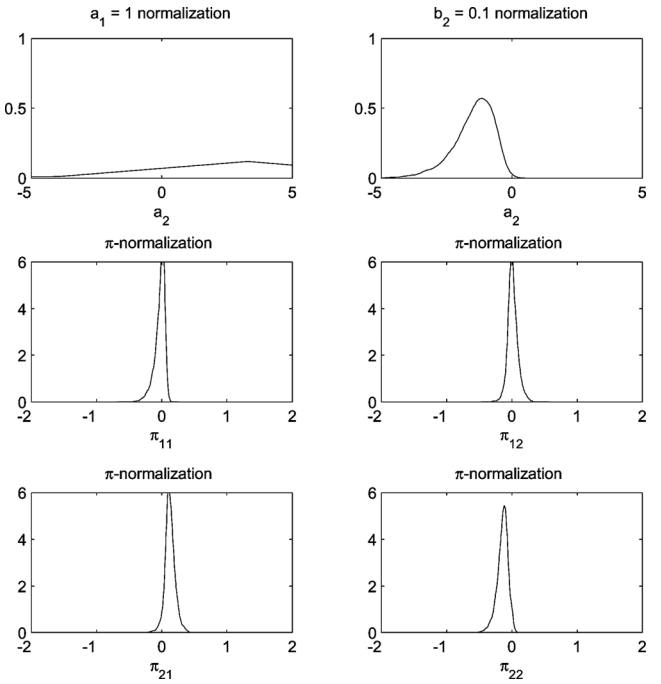

FIGURE 8 Sampling density of parameter estimates under alternative normalizations. Upper left: sampling density of $\hat{a}_2$ when a_1 is normalized at 1.0. Upper right: sampling density of $\hat{a}_2$ when b_1 is normalized at .1. Next four panels all use the π-normalization, and show sampling densities of $\hat{\pi}_{11}, \hat{\pi}_{12}, \hat{\pi}_{21}$, and $\hat{\pi}_{22}$, respectively in equation (16).

The reduced-form coefficients $\hat{\pi}_1$ and $\hat{\pi}_2$ are obtained by OLS regression of Δy_{2t} on the lags of each variable. Under the normalization $a_1 = 1$, the MLE $\hat{b}_2$ is given by $\hat{\pi}_1$ and the MLE $\hat{a}_2$ is $\hat{\pi}_2/\hat{\pi}_1$. Because there is a substantial probability of drawing a value of $\hat{\pi}_1$ near zero, the small-sample distribution of $\hat{a}_2$ is very badly behaved. By contrast, with the identification principle normalization of $b_2 = b_2^0$, the MLE's are $\hat{a}_1 = \hat{\pi}_1/b_2^0$ and $\hat{a}_2 = \hat{\pi}_2/b_2^0$. These accurately reflect the uncertainty of the OLS estimates but do not introduce any new difficulties as a result of the normalization itself.

We were able to implement the identification principle in a straightforward fashion for this example because we assumed that we knew a priori that the true value of b_1 is zero. Consider next the case where the value of b_1 is also unknown:

$$\begin{bmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} [a_1 \quad a_2] \begin{bmatrix} y_{1,t-1} \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{bmatrix}. \tag{15}$$

For this model, the normalization, $b_2 = b_2^0$ no longer satisfies the identification principle, because the allowable parameter space includes $a_1 = a_2 = 0$, at which point b_1 is unidentified.

As in the previous section, one strategy for dealing with this case is to turn to the reduced form,

$$\Delta \mathbf{y}_t = \mathbf{\Pi} y_{t-1} + \boldsymbol{\varepsilon}_t \tag{16}$$

where cointegration restricts $\mathbf{\Pi}$ to have unit rank. The parameters to be estimated in this normalization are thus not values of matrices $\mathbf{A}$ or $\mathbf{B}$ but rather the matrix $\mathbf{\Pi}$ itself. The algorithm for such estimation is described in Appendix C. Notice that this normalization satisfies the identification principle: the representation is locally identified at all points in the allowable parameter space. We generated 10,000 samples from the model with $b_1 = 0, b_2 = .1, a_1 = 1, a_2 = -1$ and calculated the maximum likelihood estimate of $\mathbf{\Pi}$ for each sample subject to restriction that $\mathbf{\Pi}$ has rank one. The resulting small-sample distributions are plotted in the bottom four panels of Figure 8. Note that, as expected, the parameter estimates are individually well-behaved and centered around the true values.

One suggestion is that the researcher simply report results in terms of this $\mathbf{\Pi}$-normalization. For example, if our data set were the first of these 10,000 samples, then the maximum likelihood estimate of $\mathbf{\Pi}$, with small-sample standard errors as calculated across the 10,000 simulated samples, is

$$\widehat{\mathbf{\Pi}} = \begin{bmatrix} \underset{(.079)}{.049} & \underset{(.078)}{-.0521} \\ \underset{(.078)}{.140} & \underset{(.085)}{-.147} \end{bmatrix}.$$

The estimated cointegrating vector could be represented identically by either row of this matrix; for example, the maximum likelihood estimates imply that

$$\underset{(.078)}{.140} y_{1t} - \underset{(.085)}{.147} y_{2t} \sim \mathrm{I}\,(0) \tag{17}$$

or that the cointegrating vector is $(1, -1.05)'$. Although $(.140, -.147)'$ and $(1, -1.05)'$ represent the identical cointegrating vector, the former is measured in units that have an objective definition, namely, .140 is the amount by which one would change one's forecast of $y_{2,t+1}$ as a result of a one-unit change of y_{1t}, and the implied t-statistic .140/.078 is a test of the null hypothesis that this forecast would not change at all.[14] By contrast, if the parameter of interest is defined to be the second coefficient a_2 in the cointegrating vector normalized as $(1, a_2)'$, the magnitude a_2

[14]Obviously these units are preferred to those that measure the effect of y_{1t} on the forecast of $y_{1,t+1}$, which effect is in fact zero in the population for this example, and a t-test of the hypothesis that it equals zero would produce a much smaller test statistic.

is inherently less straightforward to estimate and a true small-sample confidence set for this number can be quite wild, even though one has some pretty good information about the nature of the cointegrating vector itself.

There is a large literature on the Bayesian analysis of cointegration that is relevant to our analysis discussed in Koop et al. (2006). The $a_1 = 1$ normalization is the standard normalization used in the literature (e.g., Geweke, 1996). Strachan and van Dijk (2006) point out the distortion of prior beliefs associated with this normalization, resembling the discussion in Section 3. The influence of local non-identification on the likelihood and posterior density is discussed in Kleibergen and van Dijk (1994) and the convergence problem associated with an absorbing state in the Gibbs sampler of Geweke (1996) is pointed out by Kleibergen and van Dijk (1998). These problems have led to the embedding approach of working on $\mathbf{\Pi}$ via singular value decomposition (e.g., Kleibergen and van Dijk, 1998; Kleibergen and Paap, 2002) and the cointegration space approach proposed by Villani (2005, To appear). Our $\mathbf{\Pi}$-normalization is consistent with these new approaches.

Any statement about the cointegrating vector can be translated into a statement about $\mathbf{\Pi}$, the latter having the advantage that the small-sample distribution of $\widehat{\mathbf{\Pi}}$ is much better behaved than are the distributions of transformations of $\widehat{\mathbf{\Pi}}$ that are used in other normalizations. For example, in an n-variable system, one would test the null hypothesis that the first variable does not appear in the cointegrating vector through the hypothesis $\pi_{11} = \pi_{21} = \cdots = \pi_{n1} = 0$, for which a small-sample Wald test could be constructed from the sample covariance matrix of the $\widehat{\mathbf{\Pi}}$ estimates across simulated samples. One could further use the $\mathbf{\Pi}$-normalization to describe most other magnitudes of interest, such as calculating forecasts $E(\mathbf{y}_{t+j} \mid \mathbf{y}_t, \mathbf{y}_{t-1}, \ldots, \mathbf{\Pi})$ and the fraction of the forecast MSE for any horizon attributable to shocks that are within the null space of $\mathbf{\Pi}$, from which we could measure the importance of transitory versus permanent shocks at alternative forecast horizons.

6. CONCLUSIONS AND RECOMMENDATIONS FOR APPLIED RESEARCH

This paper has described some potential problems with the small-sample distributions of parameters in a wide variety of econometric models where one has imposed a seemingly innocuous normalization. These problems are related to the way in which a poor normalization is exacerbated by the local identification problem. The novel contribution of our paper is to show that a similar issue applies to the way in which normalization can determine the topography of the likelihood surface and have undesirable effects on the properties of confidence sets even when

parameters are estimated by maximum likelihood. We called attention in such settings to the loci in the parameter space along which the model is locally unidentified or the log likelihood diverges to $-\infty$, across which the interpretation of parameters necessarily changes. The problems arise whenever one mixes together parameter values across these boundaries as if they were part of a single confidence set.

Assuming that the true parameter values do not fall exactly on such a locus, this is strictly a small-sample or weak instruments problem. Under standard fixed-parameter-value asymptotics, the sampling distribution of the MLE in a classical setting, or the posterior distribution of parameters in a Bayesian setting, will have negligible probability mass on the far side of the troublesome loci. The problem that we have highlighted in this paper could be described as the potential for a poor normalization to confound the inference problems that arise when the sample is small or the identification is relatively weak.

The ideal solution to this problem is to use these loci themselves to choose a normalization, defining the boundaries of the allowable parameter space to be the loci along which the model is locally unidentified. The practical way to check whether one has accomplished this goal with a given normalization is to make sure that the model is locally identified at all interior points in the parameter space. Alternatively, if one can find a parameterization for which there is negligible probability mass associated with those regions of the parameter space that are the opposite side of such loci from the global MLE, the problems we've highlighted will be avoided as well.

For researchers who resist both of these suggestions, four other practical pieces of advice emerge from the examples investigated here. First, if one believes that normalization has made no difference in a given application, it cannot hurt to try several different normalizations to make sure that is indeed so. Second, it in any case seems good practice to plot the small-sample distributions of parameters of interest rather than simply report the mean and standard deviation. Bimodal and wide-spread distributions like those in Figure 4 or Figure 8 can be the first clue that the researcher's confidence regions are mixing together apples and oranges. Third, in Bayesian analysis, one should check whether the normalization imposed alters the information content of the prior. Finally, any researcher would do well to understand how reduced-form parameters (which typically have none of these normalization issues) are being mapped into structural parameters of interest by the normalization imposed. Such a habit can help avoid not just the problems highlighted in this paper, but should be beneficial in a number of other dimensions as well.

APPENDIX A: BAYESIAN SIMULATIONS FOR THE MIXTURE MODEL

A.1. Benchmark Simulations

Our Bayesian simulations for the i.i.d. mixture example were based on the following prior. Let $p_1 = p$ and $p_2 = 1 - p$, for which we adopt the Beta prior

$$\pi(p_1, p_2) \propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1}, \tag{A1}$$

defined over $p_1, p_2 \in [0, 1]$ with $p_1 + p_2 = 1$. Our simulations set $\alpha_1 = \alpha_2 = 1$ (a uniform prior for p). For μ_1 and μ_2 we used

$$\pi(\mu_1, \mu_2) = \varphi\left(\begin{bmatrix}\bar{\mu}_1 \\ \bar{\mu}_2\end{bmatrix}, \begin{bmatrix}\varsigma_1^2 & 0 \\ 0 & \varsigma_2^2\end{bmatrix}\right), \tag{A2}$$

where $\varphi(\mathbf{x}, \mathbf{\Omega})$ denotes the normal pdf with mean $\mathbf{x}$ and covariance matrix $\mathbf{\Omega}$ and the restrictions $\bar{\mu}_1 = \bar{\mu}_2$ and $\varsigma_1 = \varsigma_2$ are used in the text.

Denote

$$\mathbf{y} = (y_1, \ldots, y_T)', \quad \boldsymbol{\theta} = (\mu_1, \mu_2, p_1, p_2)', \quad \mathbf{s} = (s_1, \ldots, s_T)'$$

Monte Carlo draws of $\boldsymbol{\theta}$ from the marginal posterior distribution $\pi(\boldsymbol{\theta} \mid \mathbf{y})$ can be obtained from simulating samples of $\boldsymbol{\theta}$ and $\mathbf{s}$ with the following two full conditional distributions via Gibbs sampling:

$$\pi(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}), \quad \pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}).$$

It follows from the i.i.d. structure that

$$\pi(\mathbf{s} \mid \mathbf{y}, \boldsymbol{\theta}) = \prod_{t=1}^{T} \pi(s_t \mid y_t, \boldsymbol{\theta}),$$

where

$$\pi(s_t \mid y_t, \boldsymbol{\theta}) = \frac{\pi(y_t \mid s_t, \boldsymbol{\theta})\, \pi(s_t \mid \boldsymbol{\theta})}{\sum_{s_t=1}^{2} \pi(y_t \mid s_t, \boldsymbol{\theta})\, \pi(s_t \mid \boldsymbol{\theta})}, \tag{A3}$$

with

$$\pi(y_t \mid s_t, \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{(y_t - \mu_{s_t})^2}{2}\right\},$$

$$\pi(s_t \mid \theta) = \begin{cases} p_1 & s_t = 1 \\ p_2 & s_t = 2 \end{cases},$$

$$p_1 + p_2 = 1.$$

For the second conditional posterior distribution, we have

$$\pi(\boldsymbol{\theta} \mid \mathbf{y}, \mathbf{s}) = \pi(\mu_1, \mu_2 \mid \mathbf{y}, \mathbf{s}, p_1, p_2)\, \pi(p_1, p_2 \mid \mathbf{y}, \mathbf{s}).$$

Combining the prior specified in (A1) and (A2) with the likelihood function leads to

$$\begin{aligned} \pi(p_1, p_2 \mid \mathbf{y}, \mathbf{s}) &= \pi(p_1, p_2 \mid \mathbf{s}) \\ &\propto \pi(\mathbf{s} \mid p_1, p_2)\, \pi(p_1, p_2) \\ &\propto p_1^{T_1+\alpha_1-1}\, p_2^{T_2+\alpha_2-1}, \qquad \text{(A4)} \\ \pi(\mu_1, \mu_2 \mid \mathbf{y}, \mathbf{s}, p_1, p_2) &\propto \pi(\mathbf{y} \mid \mathbf{s}, p_1, p_2, \mu_1, \mu_2)\, \pi(\mu_1, \mu_2) \\ &= \varphi\left(\begin{bmatrix}\tilde{\mu}_1 \\ \tilde{\mu}_2\end{bmatrix}, \begin{bmatrix}\frac{\varsigma_1^2}{\varsigma_1^2 T_1+1} & 0 \\ 0 & \frac{\varsigma_2^2}{\varsigma_2^2 T_2+1}\end{bmatrix}\right), \qquad \text{(A5)} \end{aligned}$$

where T_k is the number of observations in state k for $k = 1, 2$ so that $T_1 + T_2 = T$ and

$$\tilde{\mu}_k = \frac{\varsigma_k^2 \sum_{t=k(1)}^{k(T_k)} y_t + \bar{\mu}_k}{\varsigma_k^2 T_k + 1}, \; s_{k(q)} = k \quad \text{for } q = 1, \ldots, T_k, \;\; k = 1, 2.$$

The posterior density (A4) is of Beta form and (A5) is of Gaussian form; thus, sampling from these distributions is straightforward.

A.2. Truncated Gaussian Prior

The truncated Gaussian prior used in the text has the form

$$\pi(\mu_1, \mu_2) = \pi(\mu_1)\pi(\mu_2 \mid \mu_1), \qquad \text{(A6)}$$

where $\pi(\mu_2 \mid \mu_1)$ is given by (3). Replacing the symmetric prior (A2) with the truncated prior (A6) leads to the following posterior pdf of μ_1 and μ_2:

$$\pi(\mu_1, \mu_2 \mid \mathbf{y}, \mathbf{s}, p_1, p_2) = \frac{1}{\Phi\left(\frac{\mu_1 - \tilde{\mu}_2}{\varsigma_2}\right)} \varphi\left(\begin{bmatrix}\tilde{\mu}_1 \\ \tilde{\mu}_2\end{bmatrix}, \begin{bmatrix}\frac{\varsigma_1^2}{\varsigma_1^2 T_1+1} & 0 \\ 0 & \frac{\varsigma_2^2}{\varsigma_2^2 T_2+1}\end{bmatrix}\right) \qquad \text{(A7)}$$

if $\mu_2 \leq \mu_1$ and zero otherwise.

Because $\bar{\mu}_2 \neq \tilde{\mu}_2$ and $\varsigma^2 \neq \frac{\varsigma_2^2}{\varsigma_2^2 T_2+1}$, the conditional posterior pdf (A7) is not of any standard form. To sample from (A7), we use a Metropolis

algorithm (e.g., Chib and Greenberg, 1996) with the transition pdf of μ' conditional on the jth draw $\mu^{(j)}$ given by

$$q(\mu^{(j)}, \mu' \mid \mathbf{y}, \mathbf{s}, p_1, p_2) = \varphi\left(\begin{bmatrix}\mu_1^{(j)}\\ \mu_2^{(j)}\end{bmatrix}, \begin{bmatrix}\frac{s_1^2}{s_1^2 T_1+1} & 0\\ 0 & \frac{s_2^2}{s_2^2 T_2+1}\end{bmatrix}\right), \tag{A8}$$

where c is a scaling factor to be adjusted to maintain an optimal acceptance ratio (e.g., between 25% and 40%). Given the previous posterior draw $\mu^{(j)}$, the algorithm sets $\mu^{(j+1)} = \mu'$ with acceptance probability[15]

$$\min\left\{1, \frac{\pi(\mu' \mid \mathbf{y}, \mathbf{s}, p_1, p_2)}{\pi(\mu^{(j)} \mid \mathbf{y}, \mathbf{s}, p_1, p_2)}\right\} \quad \text{if } \mu_2' < \mu_1';$$

otherwise, the algorithm sets $\mu^{(j+1)} = \mu^{(j)}$.[16]

APPENDIX B: SIMULATING ALGORITHM FOR THE VAR MODEL

We describe our algorithm for simulating VAR distributions in terms of the following representation, obtained by premultiplying (4) by $\mathbf{D}^{-1}$ and transposing,

$$\mathbf{y}_t'\mathbf{A}_0 = \mathbf{c} + \mathbf{y}_{t-1}'\mathbf{A}_1 + \mathbf{y}_{t-2}'\mathbf{A}_2 + \cdots + \mathbf{y}_{t-p}'\mathbf{A}_p + \boldsymbol{\varepsilon}_t' \tag{A9}$$

where $\mathbf{A}_j = \mathbf{B}_j'\mathbf{D}^{-1}$ for $j = 0, 1, \ldots, p$ and $\boldsymbol{\varepsilon}_t = \mathbf{D}^{-1}\mathbf{u}_t$ so that $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \mathbf{I}_n$. All simulations were done using the Gibbs sampler for structural VARs described in Waggoner and Zha (2003a). This technique samples from the posterior distribution associated with the specification given by (A9). A flat prior was used to obtain draws of $\mathbf{A}_0, \mathbf{A}_1, \ldots, \mathbf{A}_p$ and then these parameters were transformed into the other specifications used in this paper. Because these transformations are non-linear, the Jacobian is non-trivial, and the resulting draws for the alternate specifications will have diffuse, as opposed to flat, priors. In the case of the β and η normalizations, the likelihood is not proper[17], so the posterior will not be proper unless some sort of prior is imposed. A direct computation reveals that the Jacobian involves only the variance terms and it tends to favor smaller values for the variance. The prior on the parameters of interest γ, h, and β or η will be flat.

[15]Note from (A8) that $q(\mu, \mu') = q(\mu', \mu)$, allowing us to use the Metropolis as opposed to the Metropolis–Hastings algorithm.

[16]If the random value $\mu_1^* = \mu_1'$ generated from $q(\mu^{(j)}, \mu' \mid \mathbf{y}, \mathbf{s}, p_1, p_2)$ or $\mu_1^* = \mu_1^{(j)}$ results in a numerical underflow when $\Phi\left(\frac{\mu_1^* - \bar{\mu}_2}{s_2}\right)$ is calculated, we could always set $\mu^{(j+1)} = \mu'$ as an approximation to a draw from the Metropolis algorithm. In our simulations, however, such an instance did not occur.

[17]See Sims and Zha (1994) for a discussion of this result.

The likelihood for the π-normalization is proper and so in theory one could impose the flat prior for this case. Though the Jacobian in this case is difficult to interpret, we note that the π-normalization is similar to the reduced form specification. The technique used in this paper, applied to the reduced form specification, would be equivalent to using a flat prior on the reduced form, but with the sample size increased.

APPENDIX C: MLE FOR THE COINTEGRATION MODEL

Maximum likelihood estimation of (16) can be found using the Anderson (1984)–Johansen (1988) procedure, as described in (Hamilton, 1994, p. 637). Specifically, let $\hat{\mathbf{\Sigma}}_{\mathbf{vv}} = T^{-1}\sum_{t=1}^{T}\mathbf{y}_{t-1}\mathbf{y}_{t-1}'$, $\hat{\mathbf{\Sigma}}_{\mathbf{uu}} = T^{-1}\sum_{t=1}^{T}\Delta\mathbf{y}_t\Delta\mathbf{y}_t'$, $\hat{\mathbf{\Sigma}}_{\mathbf{uv}} = T^{-1}\sum_{t=1}^{T}\Delta\mathbf{y}_t\mathbf{y}_{t-1}'$, and $\hat{\mathbf{P}} = \hat{\mathbf{\Sigma}}_{\mathbf{vv}}^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{uv}}'\hat{\mathbf{\Sigma}}_{\mathbf{uu}}^{-1}\hat{\mathbf{\Sigma}}_{\mathbf{uv}}$. Find $\tilde{\mathbf{a}}_1$, the eigenvector of $\hat{\mathbf{P}}$ associated with the biggest eigenvalue and construct $\hat{\mathbf{a}}_1 = \tilde{\mathbf{a}}_1/\sqrt{\tilde{\mathbf{a}}_1'\hat{\mathbf{\Sigma}}_{\mathbf{vv}}\tilde{\mathbf{a}}_1}$. The MLE is then $\hat{\mathbf{\Pi}} = \hat{\mathbf{\Sigma}}_{\mathbf{uv}}\hat{\mathbf{a}}_1\hat{\mathbf{a}}_1'$.

ACKNOWLEDGMENT

We thank two referees, Søren Johansen, and Adrian Pagan for helpful and critical comments. This research was supported by the National Science Foundation under Grant No. SES-0215754.

The views expressed here are not necessarily those of the Federal Reserve System or the Federal Reserve Bank of Atlanta. Computer code used in this study can be downloaded free of charge from ftp://weber.ucsd.edu/pub/jhamilto/hwz.zip.

REFERENCES

Aitkin, M., Rubin, D. B. (1985). Estimation and hypothesis testing in finite mixture models. *J. Royal Statist. Soc. Series B* 47:67–75.

Anderson, T. (1984). *An Introduction to Multivariate Statistical Analysis.* 2nd ed. New York: John Wiley & Sons.

Anderson, T. W. (2005). Origins of the limited information maximum likelihood and two-stage least squares estimators. *J. Econometrics* 127:1–16.

Celeux, G., Hurn, M., Robert, C. P. (2000). Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Statist. Assoc.* 95:957–970.

Chao, J. C., Phillips, P. C. (2005). Jeffreys prior analysis of the simultaneous equations model in the case with $n+1$ endogenous variables. *J. Econometrics* 111(2):251–283.

Chao, J. C., Swanson, N. R. (2005). Consistent estimation with a large number of weak instruments. *Econometrica* 73(5):1673–1692.

Chen, A. M., Minn Lu, H., Hecht-Nielsen, R. (1993). On the geometry of feedforward neural network error surfaces. *Neural Computation* 5(6):910–927.

Chib, S., Greenberg, E. (1996). Markov chain monte carlo simulation methods in econometrics. *Econometric Theory* 12:409–431.

Drèze, J. H. (1976). Bayesian limited information analysis of the simultaneous equations. *Econometrica* 44:1045–1075.

Drèze, J. H., Morales, J. A. (1976). Bayesian full information analysis of simultaneous equations. *J. Amer. Statist. Assoc.* 71(356):919–923.
Drèze, J. H., Richard, J. F. (1983). Bayesian analysis of simultaneous equation systems. In: Griliches, Z., Intriligator, M., eds. Vol. 1. *Handbook of Econometrics*. Amsterdam: North-Holland.
Forchini, G., Hillier, G. (2003). Conditional inference for possibly unidentified structural equations. *Econometric Theory* 19:707–743.
Frühwirth-Schnatter, S. (2001). Markov chain monte carlo estimation of classical and dynamic switching and mixture models. *J. Amer. Statist. Assoc.* 96(453):194–209.
Geweke, J. (1996). Bayesian reduced rank regression in econometrics. *J. Econometrics* 75:121–146.
Hahn, J., Hausman, J. (2002). A new specification test for the validity of instrumental variables. *Econometrica* 70:163–189.
Hamilton, J. D. (1989). A new approach to the economic analysis of nonstationary time series and the business cycle. *Econometrica* 57(2):357–384.
Hamilton, J. D. (1994). *Times Series Analysis*. Princeton, NJ: Princeton University Press.
Hauck W. W. Jr., Donner, A. (1977). Wald's test as applied to hypotheses in logit analysis. *J. Amer. Statist. Assoc.* 72:163–189.
Johansen, S. (1988). Statistical analysis of cointegration vectors. *J. Economic Dynamics and Control* 12:231–254.
Kleibergen, F., van Dijk, H. K. (1994). On the shape of the likelihood/posterior in cointegration models. *Econometric Theory* 10:514–551.
Kleibergen, F., van Dijk, H. K. (1998). Bayesian simultaneous equations analysis using reduced rank structures. *Econometric Theory* 14:701–743.
Kleibergen, F., Paap, R. (2002). Priors, posteriors, and bayes factors for a bayesian analysis of cointegration. *J. Econometrics* 111:223–249.
Koop, G., Strachan, R., van Dijk, H., Villani, M. (2006). Bayesian approaches to cointegration. In: Mills, T. C., Patterson, K., eds. *The Palgrave Handbook of Econometrics: Theoretical Econometrics*. Vol. 1. Basingstoke, England: Palgrave Macmillan, Chapter 25.
Koopmans, T. C. (1953). Identification problems in economic model construction. In: Hood, W. C., Koopmans, T. C., eds. *Studies in Econometric Method*. New York, New York: Wiley.
Leeper, E. M., Sims, C. A., Zha, T. (1996). What does monetary policy do? *Brookings Papers on Economic Activity* 2:1–78.
Lenk, P. J., DeSarbo, W. S. (2000). Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika* 65:93–119.
Litterman, R. B. (1986). Forecasting with bayesian vector autoregressions—five years of experience. *J. Business and Economic Statistics* 4:25–38.
Maddala, G. (1983). *Limited-Dependent and Qualitative Variables in Econometrics*. Cambridge: Cambridge University Press.
Manski, C. F. (1988). Identification of binary response models. *Journal of the American Statistical Association* 83:729–738.
Otrok, C., Whiteman, C. H. (1998). Bayesian leading indicators: measuring and predicting economic conditions in Iowa. *International Economic Review* 39:997–1014.
Pagan, A. R., Robertson, J. (1997). GMM and its problems. Working Paper, Australian National University.
Phillips, P. C. B. (1994). Some exact distribution theory for maximum likelihood estimators of cointegrating coefficients in error correction models. *Econometrica* 62:73–93.
Rothenberg, T. J. (1971). Identification in parametric models. *Econometrica* 39:577–591.
Ruben, H. (1954). On the moments of order statistics in samples from normal populations. *Biometrika* 41:200–227.
Rüger, S. M., Ossen, A. (1996). Clustering in weight space of feedforward nets. In: von der Malsburg, C., von Seelen, W., Vorbrüggen, J. C., Sendhoff, B., eds. *Proceedings of the International Conference on Artificial Neural Networks* (ICANN-96, Bochum). Vol. 1125. Bochum: Springer, pp. 83–88.
Sims, C. A. (2005). Dummy observation priors revisited. Manuscript, Princeton University.
Sims, C. A., Zha, T. (1994). Error bands for impulse responses. Cowles Foundation Discussion Paper No. 1085.
Sims, C. A., Zha, T. (1998). Bayesian methods for dynamic multivariate models. *International Economic Review* 39(4):949–968.

Sims, C. A., Zha, T. (1999). Error bands for impulse responses. *Econometrica* 67(5):1113–1155.
Smith, P. A., Summers, P. M. (2003). Identification and normalization in markov switching models of business cycles. Working paper, University of Melbourne.
Staiger, D., Stock, J. H. (1997). Instrumental variables regression with weak instrruments. *Econometrica* 65:557–586.
Stephens, M. (2000). Dealing with label switching in mixture models. *J. Royal Statistical Society Series B* 62:795–809.
Stock, J. H., Yogo, M. (2005). Testing for weak instruments in linear IV regression. In: Andrews, D. W., Stock, J. H., eds. *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*. Cambridge, England: Cambridge University Press.
Stock, J. H., Wright, J. H., Yogo, M. (2002). A survey of weak instruments and weak identification in generalized method of moments. *J. Business and Economic Statistics* 20:518–529.
Strachan, R., van Dijk, H. K. (2006). Model uncertainty and bayesian model averaging in vector autoregressive processes. University of Leicester Working Paper No. 06/5.
Villani, M. (2005). Bayesian reference analysis of cointegration. *Econometric Theory* 21:326–357.
Villani, M. Bayesian point estimation of the cointegration space. To appear in *J. Econometrics* 134:645–664.
Waggoner, D. F., Zha. T. (2003a). A gibbs sampler for structural vector autoregressions. *J. Economic Dynamics and Control* 28(2):349–366.
Waggoner, D. F., Zha, T. (2003b). Likelihood preserving normalization in multiple equation models. *J. Econometrics* 114(2):329–347.
Yogo, M. (2004). Estimating the elasticity of intertemporal substitution when instruments are weak. *Review of Economics and Statistics* 86:797–810.