FEDERAL RESERVE BANK *of* ATLANTA

Normalization in Econometrics

James D. Hamilton, Daniel F. Waggoner, and Tao Zha

Working Paper 2004-13
June 2004

WORKING PAPER SERIES

FEDERAL RESERVE BANK *of* ATLANTA WORKING PAPER SERIES

Normalization in Econometrics

James D. Hamilton, Daniel F. Waggoner, and Tao Zha

Working Paper 2004-13
June 2004

Abstract: The issue of normalization arises whenever two different values for a vector of unknown parameters imply the identical economic model. A normalization does not just imply a rule for selecting which point, among equivalent ones, to call the maximum likelihood estimator (MLE). It 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, confidence intervals that are disjoint, and very misleading characterizations of the true statistical uncertainty. This paper introduces the 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. The authors illustrate these issues with examples taken from mixture models, structural VARs, and cointegration.

JEL classification: C1

Key words: normalization, mixture distributions, vector autoregressions, cointegration, regime switching, numerical Bayesian methods, small sample distributions, weak identification

This research was supported by the National Science Foundation under Grant No. SES-0215754. Computer code used in this study can be downloaded free of charge from ftp://weber.ucsd.edu/pub/jhamilto/hwz.zip. The views expressed here are the authors' and not necessarily those of the Federal Reserve Bank of Atlanta or the Federal Reserve System. Any remaining errors are the authors' responsibility.

Please address questions regarding content to James D. Hamilton, Department of Economics, 0508, University of California, San Diego, La Jolla, CA 92093-0508, jhamilton@ucsd.edu; Daniel F. Waggoner, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street N.E., Atlanta, GA 30309-4470, 404-498-8278, daniel.f.waggoner@atl.frb.org; or Tao Zha, Research Department, Federal Reserve Bank of Atlanta, 1000 Peachtree Street N.E., Atlanta, GA 30309-4470, 404-498-8353, tao.zha@atl.frb.org.

1 Introduction.

The issue of normalization arises whenever two different values for a vector of unknown parameters imply the identical economic model. Although one's initial reaction might be that it makes no difference how one resolves this ambiguity, a host of counterexamples establish that it can matter a great deal.

Certainly it is well understood that a given nonlinear restriction $g(\boldsymbol{\theta}) = 0$ can be written in an infinite number of ways, for example, as $g^*(\boldsymbol{\theta}) = \theta_1 g(\boldsymbol{\theta}) = 0$ for θ_1 the first element of the vector $\boldsymbol{\theta}$. Any approach derived from the asymptotic Taylor approximation $g(\hat{\boldsymbol{\theta}}) \simeq g(\boldsymbol{\theta}) + \nabla g(\boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ can of course give different results for different functions $g(.)$ in a given finite sample. Indeed, one can obtain any result one likes for a given sample (accept or reject the null hypothesis) by judicious choice of $g(.)$ in formulating a Wald test of a nonlinear restriction (e.g., Gregory and Veall, 1985).

Likewise, a model implying $E[\mathbf{g}(\boldsymbol{\theta}; \mathbf{y}_t)] = \mathbf{0}$ can be written an infinite number of ways. For example, the instrumental variable orthogonality condition $E[(y_t - \beta z_t)\mathbf{x}_t] = \mathbf{0}$ can equivalently be written as the reverse regression $E[(z_t - \alpha y_t)\mathbf{x}_t] = \mathbf{0}$ for $\alpha = 1/\beta$, though for $\mathbf{x}_t$ a vector, the GMM estimates fail to satisfy $\hat{\alpha} = 1/\hat{\beta}$ in a finite sample. The appropriate way to normalize such GMM conditions and their relation to the weak instrument problem has been the focus of much recent research, including Pagan and Robertson (1997), Hahn and Hausman (2002), and Yogo (2003).

Less widely appreciated is the fact that normalization can also materially affect the conclusions one draws with likelihood-based methods. 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 themselves are invariant with respect to a reparameterization, one might think that how one normalizes would be of no material relevance in such cases. If one were interested in reporting only the maximum likelihood estimate 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 $\boldsymbol{\theta}_1$. 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. 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. Geweke (1996) and particularly Kleibergen and Paap (2002) noted the importance of normalization for Bayesian analysis of cointegrated systems. The "label-switching" problem for mixture models has an extensive statistics literature discussed below. Waggoner and Zha (2003a) discussed normalization in structural VARs, documenting with a very well-known model that, if one calculates standard errors for impulse-response coefficients using the algorithms that researchers have been relying on for twenty years, the resulting 95% confidence bounds are 60% wider than they should be under a better normalization. Waggoner and Zha suggested a principle for normalizing structural VARs, which they called the "likelihood principle," that avoids these problems.

To our knowledge, all previous discussions have dealt with normalization in terms of the specific issues arising in a particular class of models, 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, which we call the "identification principle," which turns out to include Waggoner and Zha's likelihood principle as a special case.

We find it helpful to introduce the key issues in Section 2 with a transparently 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 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 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, Lu, and Hecht-Nielsen, 1993; Rüger and Ossen, 1996). Section 6 summarizes our practical recommendations for applied research in any setting requiring normalization.

2 Normalization and the identification principle.

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. $N(0,1)$. Figure 1 plots the log likelihood for a sample of size $T = 50$,

$$\log f(y_1, ..., y_T; \sigma) = -(T/2)\log(2\pi) - (T/2)\log(\sigma^2) - \sum_{t=1}^{T} y_t^2/(2\sigma^2),$$

as a function of σ, with the true $\sigma_0 = 1$. The likelihood function is of course symmetric, 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 restrict $\sigma > 0$. 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 $\pi(\sigma|\mathbf{y})$ 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 $\pi(\sigma|\mathbf{y})$ 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. 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 this 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

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

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.

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, unless there are discontinuities in the likelihood surface, there must be loci in $\Re^k$ along which the structure is locally unidentified as well. These loci characterize the boundaries along which the interpretation of parameters is fundamentally ambiguous and across which the interpretation of parameters necessarily changes. 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. The check of a candidate normalization A is thus to make sure that the structure is locally identified at all interior points of A. We describe this as choosing a normalization according to the *identification principle.*

The loci along which the observationally equivalent structures ($\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$) merge together typically take one of two forms. If elements of the information matrix,

$$\Im(\boldsymbol{\theta}) = -E\left(\frac{\partial^2 f(\mathbf{y};\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\,\partial\boldsymbol{\theta}'}\right) \tag{1}$$

exist and are everywhere continuous, then these loci are characterized by the points in $\Re^k$ for which the information matrix becomes singular; see Rothenberg (1971). A second case occurs when the log likelihood diverges to $-\infty$ when approached from either side of the locus.

Figure 1 represents an example of the second case. Here, since $k = 1$, the locus is simply a point in $\Re^1$, namely, $\sigma = 0$. Using this locus as the boundary for A means defining A by the condition $\sigma > 0$, the common sense normalization for this transparent example.

In their analysis of structural VAR's, Waggoner and Zha (2003a) suggest using an algorithm that ensures that any candidate value $\boldsymbol{\theta}$ satisfy the condition $f(\mathbf{y};\boldsymbol{\theta}^*) > 0$ for $\boldsymbol{\theta}^* = s\boldsymbol{\theta}+(1-s)\hat{\boldsymbol{\theta}}$ for all $s \in [0,1]$ and $\hat{\boldsymbol{\theta}}$ the MLE, which they refer to as the *likelihood principle* for normalization. This condition prevents $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}$ from falling on opposite sides of any locus along which the log likelihood is $-\infty$, and thus has the consequence of using these loci to determine the boundaries of A. Thus the likelihood principle is a special case of the identification principle.

The following sections illustrate these ideas in a number of different settings.

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 following 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

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

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, Hurn, and Robert (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 (2) 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 $f(.)$. 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, deliberately introducing jumps in the simulation chain to make 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, Hurn, and Robert, 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 $f(.)$, 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 $p > 0.5$, as in Aitkin and Rubin (1985) or Lenk and DeSarbo (2000). However, according to the identification principle discussed in the introduction, this is not a satisfactory solution to the normalization problem. This is because even if one restricts $p > 0.5$, 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 $p = 0.8$. Figure 2 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 at $\mu_1 = -1$, $\mu_2 = 1$, and $p = 0.2$. In the middle right panel ($p = 0.5$), points above the 45^o line are the mirror image of those below. The proposed normalization ($p > 0.5$) restricts the space to the first 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^o line that is an echo of the exact symmetry of the $p = 0.5$ panel. Conditional on any value of p, the normalization $p > 0.5$ 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 $p > 0.5$ 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. In contrast to the first normalization, the normalization $\mu_1 > \mu_2$ satisfies the identification principle– $\boldsymbol{\theta}$ is locally identified at all points interior to this region. 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 $p^{(i)} > 0.5$, but used $(\mu_2^{(i)}, \mu_1^{(i)}, 1-p^{(i)})'$ otherwise. We ran the Gibbs sampler for 5500 iterations on each sample, with parameter values initialized

[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, Hurn and Rober (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, Hurn and Robert (2000) and Stephens (2000) proposed a decision-theoretic framework whose relation to our approach is commented on below.

from the prior, discarded the first 500 iterations, and interpreted the last 5000 iterations as draws from the posterior distribution of parameters for that sample. We repeated this process on 1000 different samples each of size $T = 50$. The Bayesian posterior densities (regarding these as 5,000,000 draws from a single distribution) are plotted in Figure 3.[2] The distribution of μ_1 is downward biased as a result of a bulge in the distribution, which represents μ_2 estimates that get labelled as μ_1. More noticeable is the upward bias for μ_2 introduced from the same label switching. And although the average value of p is centered around the true value of 0.8, this results almost by force from the normalization $p > 0.5$; it is not clear that the information in the sample has been used in any meaningful way to refine the estimate of p.

Figure 4 presents posterior distributions for the $\mu_1 > \mu_2$ normalization. The distributions for μ_1 and μ_2 are both much more reasonable. The distribution for μ_2 is still substantially spread out and upward biased, though there is simply little information in the data about this parameter, as a typical sample contains only ten observations from distribution 2. The distribution of p has its peak near the true value of 0.8, but also is somewhat spread out and has significant mass for small values. Evidently, there are still a few samples in which label switching has taken place with this normalization, despite the fact that it satisfies our

[2] These are nonparametric densities calculated with a triangular kernel (e.g., Silverman, 1992, pp. 27 and 43):

$$\hat{f}_\theta(t) = \frac{1}{hI}\sum_{i=1}^{I} \max\left(0, 1 - h^{-1}|\theta^{(i)} - t|\right)$$

where $\theta^{(i)}$ is the ith Monte Carlo draw of a given parameter θ, $I = 5,000,000$ is the number of Monte Carlo draws, and $\hat{f}_\theta(t)$ is our estimate of the value of the density of the parameter θ evaluated at the point $\theta = t$. The bandwidth h was taken to be 0.01.

identification principle.

Identification as defined by Rothenberg (1971) is a population property– parameters $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are observationally distinguishable if there exists *some* potentially observable sample $(y_1, \ldots, y_T)$ for which the values would imply different values for the likelihood. For local identification, the appropriate measure is therefore based on the population information matrix (1). However, even though a structure may be theoretically identified, the identification can be weak, in the sense that there is very little information in a particular observed sample that allows us to distinguish between related points. For example, consider those samples in the above simulations in which very few observations were generated from distribution 2. The posterior distribution for μ_2 for these samples is very flat, and a large value of μ_2 is likely to be drawn and labeled as representing state 1 by the rule $\mu_1 > \mu_2$. This makes the inference of μ_1 and p contaminated and distorted. For this reason, it may be helpful to consider another version of the identification principle based on the sample hessian,

$$\mathbf{H}(\theta) = -\sum_{t=1}^{T} \frac{\partial^2 \ \log f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta} \ \partial \boldsymbol{\theta}'}, \tag{3}$$

evaluated at the MLE $\hat{\boldsymbol{\theta}}$.

To motivate this alternative implementation of the identification principle, let $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ represent two parameter values that imply the identical population probability law, so that $f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1; \boldsymbol{\theta}_1) = f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1; \boldsymbol{\theta}_2)$ for all possible realizations. Suppose that for the observed sample, in a local neighborhood around the maximum likelihood estimate $\hat{\boldsymbol{\theta}} \in \Re^k$, iso-likelihood contours of the log likelihood were exact spheroids in $\Re^k$.[3]

Then if $\boldsymbol{\theta}_1$ is closer in Euclidean distance to $\hat{\boldsymbol{\theta}}$ than is $\boldsymbol{\theta}_2$, it must be the case that the line segment connecting $\hat{\boldsymbol{\theta}}$ with $\boldsymbol{\theta}_2$ crosses a locus along which $\boldsymbol{\theta}$ is locally unidentified. Accordingly, if for any simulated $\boldsymbol{\theta}^{(i)}$ we always selected the permutation that is closest in Euclidean distance to $\hat{\boldsymbol{\theta}}$, we would implicitly be using the identification principle to specify the allowable parameter space.

This procedure would work as long as the iso-likelihood contours between $\hat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_1$ (if $\boldsymbol{\theta}_1$ is the value selected by this approach) are all exact spheroids. But note that we can always reparameterize the likelihood in terms of $\boldsymbol{\lambda} = \mathbf{K}\boldsymbol{\theta}$ so as to to make the likelihood contours in terms of $\boldsymbol{\lambda}$ approximate spheroids in the vicinity of $\hat{\boldsymbol{\theta}}$, by choosing $\mathbf{K}$ to be the Cholesky factor of $\mathbf{H}(\hat{\boldsymbol{\theta}})$.[4] Choosing the value of $\boldsymbol{\lambda}$ that is closest to $\hat{\boldsymbol{\lambda}}$ is equivalent to choosing the value of $\boldsymbol{\theta}$ that minimizes

$$(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathbf{H}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}). \tag{4}$$

Expression (4) will be recognized as the Wald statistic for testing the null hypothesis that $\boldsymbol{\theta}$ is the true parameter value.

The procedure to implement this idea works as follows. After simulating the ith draw

[3] Obviously the log likelihood cannot globally be perfect spheroids since the equivalence of $f(\mathsf{y}_t|\mathsf{y}_{t-1}, \mathsf{y}_{t-2}, ..., \mathsf{y}_1, \boldsymbol{\theta}_1)$ with $f(\mathsf{y}_t|\mathsf{y}_{t-1}, \mathsf{y}_{t-2}, ..., \mathsf{y}_1, \boldsymbol{\theta}_2)$ implies there must be a saddle when one gets far enough away from $\hat{\boldsymbol{\theta}}$.

[4] Let $\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log f(\mathsf{y}_t|\mathsf{y}_{t-1}, \mathsf{y}_{t-2}, ..., \mathsf{y}_1; \boldsymbol{\theta})$ be the log likelihood. To a second-order Taylor approximation,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &\cong \mathcal{L}(\hat{\boldsymbol{\theta}}) - (1/2)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathsf{H}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= \mathcal{L}(\hat{\boldsymbol{\theta}}) - (1/2)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})'\mathsf{K}'\mathsf{K}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \\ &= \mathcal{L}(\mathsf{K}^{-1}\hat{\boldsymbol{\lambda}}) - (1/2)(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})'(\boldsymbol{\lambda} - \hat{\boldsymbol{\lambda}})\end{aligned}$$

whose contours as functions of $\boldsymbol{\lambda}$ are spheroids.

$\boldsymbol{\theta}^{(i)}$, calculate all the observationally equivalent permutations of this draw $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \ldots, \boldsymbol{\theta}_M^{(i)})$. For each $\boldsymbol{\theta}_m^{(i)}$, calculate $(\boldsymbol{\theta}_m^{(i)} - \hat{\boldsymbol{\theta}})' \mathbf{H}(\hat{\boldsymbol{\theta}})(\boldsymbol{\theta}_m^{(i)} - \hat{\boldsymbol{\theta}})$ as if testing the null hypothesis that $\boldsymbol{\theta} = \boldsymbol{\theta}_m^{(i)}$ where $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate. The actual value for $\boldsymbol{\theta}_m^{(i)}$ that is used for $\boldsymbol{\theta}^{(i)}$ is the one with the minimum Wald statistic. We will refer to this as the Wald normalization.

In practice we have found that the algorithm is slightly more robust when we replace the second-derivative estimate in (3) with its outer-product analog:

$$\hat{\mathbf{H}} = \sum_{t=1}^{T} \left[\frac{\partial \ \ln f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right] \left[\frac{\partial \ \ln f(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_1; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right]' \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}} . \qquad (5)$$

Note that the Wald normalization is related to the decision-theoretic normalizations proposed by Celeux, Hurn and Robert (2000) and Stephens (2000). They suggested that the ideal normalization should minimize the posterior expected loss function. For example, in Stephens's formulation, one selects the m_i for which the loss function $\mathsf{L}_0(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}_{m_{\mathsf{i}}}^{(i)})$ is smallest. Stephens proposed implementing this by iterating on a zig-zag algorithm, first taking the normalization for each draw (a specification $m_1, m_2, ..., m_N$ for N the number of Monte Carlo draws) as given and choosing $\tilde{\boldsymbol{\theta}}$ so as to minimize $N^{-1} \sum_{i=1}^{N} \mathsf{L}_0(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}_{m_{\mathsf{i}}}^{(i)})$, and then taking $\tilde{\boldsymbol{\theta}}$ as given and selecting a new normalization m_i for the ith draw so as to minimize $\mathsf{L}_0(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}_{m_{\mathsf{i}}}^{(i)})$. Our procedure would thus correspond to the decision-theoretic optimal normalization if the loss function were taken to be $\mathsf{L}_0(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}) = (\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{H}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$ and we were to adopt a Stephens zig-zag iteration, replacing the MLE $\hat{\boldsymbol{\theta}}$ at each zag with that iteration's Bayesian posterior estimate (the minimizer of $N^{-1} \sum_{i=1}^{N} \mathsf{L}(\tilde{\boldsymbol{\theta}}; \boldsymbol{\theta}_{m_{\mathsf{i}}}^{(i)})$). Our specification is also closely related to the Celeux, Hurn and Robert's loss function that minimizes $(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})' \mathbf{S}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}})$ for $\mathbf{S}(\tilde{\boldsymbol{\theta}})$ a diagonal matrix containing reciprocals of the variance of elements of $\boldsymbol{\theta}$ across

Monte Carlo draws.

Figure 5 displays the posterior densities for the Wald normalization. These offer a clear improvement over the $p > 0.5$ normalization, and do better at describing the density of p than does the $\mu_1 > \mu_2$ normalization. Unfortunately, the Wald normalization seems to do slightly worse at describing the distributions of μ_1 and μ_2 than does the $\mu_1 > \mu_2$ normalization.

We repeated the above calculations for samples of size $T = 10$, 20, 50, and 100. For the nth generated sample, we calculated the difference between the posterior mean $E(\boldsymbol{\theta}|\mathbf{y}^{(n)})$ 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 Figure 6. 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. The $p > 0.5$ and Wald normalizations do substantially better than the $\mu_1 > \mu_2$ normalization in terms of estimating p. Curiously, though, the MSE for the $p > 0.5$ normalization deteriorates as sample size increases.

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. Figure 7 reports these 90% coverage probabilities for the three normalizations. Although we have seen that the $p > 0.5$ and Wald normalizations produce substantially better point estimates of p, they significantly distort the distribution. The $\mu_1 > \mu_2$ normalization, despite its poorer point estimate, would produce a more accurately sized test of the null hypothesis $p = p_0$. It also produces the most accurately sized test of the hypotheses $\mu_1 = \mu_{10}$ or $\mu_2 = \mu_{20}$ for large samples.

The superior point estimate of the parameter p that is obtained with the $p > 0.5$ normalization in part results from the interaction between the normalization rule and the prior. Note that the uniform prior for p implies that with no normalization (or with a normalization based solely on $\mu_1 > \mu_2$), the prior expectation of p is 0.5. However, when a uniform prior is put together with the $p > 0.5$ normalization, this implies a prior expectation of 0.75.[5] Given the true value of $p = 0.8$ used in the simulations, the normalization turns what was originally a vague prior into quite a useful description of the truth.

The normalization $\mu_1 > \mu_2$ similarly interacts with the prior for μ in this case to substantially improve the accuracy of the prior information. If the prior is

$$\begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix} \sim N\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \varsigma^2 & 0 \\ 0 & \varsigma^2 \end{bmatrix} \right), \tag{6}$$

then $E(\mu_1^* = \max\{\mu_1, \mu_2\}) = \varsigma/\sqrt{\pi}$. [6] 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

[5] If $p \sim N(U(0, 1)$ and $p^* = \max\{p, 1 - p\}$, then $E(p^*) = 0.75$.

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

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.[7] Likewise the prior is also assigning a 50% probability that $\mu_2 < -5$, when the event $y_t < -5$ is also virtually impossible.

Figure 8 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 top panel of Figure 8 is identical to the solid line in the top panel of Figure 6, 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 Figure 8). 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 either of the other normalizations 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 (6), it seems better to employ a truncated Gaussian prior, where $\mu_1 \sim N(\overline{\mu}_1, \varsigma_1^2)$ and

$$\pi(\mu_2|\mu_1) = \begin{cases} \frac{1}{\Phi[(\mu_1-\overline{\mu}_2)/\varsigma_2]\sqrt{2\pi}\varsigma_2} \exp\left(\frac{-(\mu_2-\overline{\mu}_2)^2}{2\varsigma_2^2}\right) & \text{if } \mu_2 \leq \mu_1 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

[7] The probability that a variable drawn from the distribution with the larger mean ($N(1, 1)$) exceeds 5.5 is 0.00000340.

for $\Phi(z) = \text{Prob}(Z \leq z)$ for $Z \sim N(0,1)$. Here $\overline{\mu}_2$ 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 $\overline{\mu}_1 = \overline{\mu}_2 = 0$ and $\varsigma_1^2 = \varsigma_2^2 = 5$. When $\mu_1 = 0$, for example, this prior implies an expected value for μ_2 of $\overline{\mu}_2 + \varsigma_2 M_2 = -1.78$ where $M_2 = -\phi(c_2)/\Phi(c_2) = -0.7979$ with $c_2 = (\mu_1 - \overline{\mu}_2)/\varsigma_2 = 0$ and a variance for μ_2 of $\varsigma_2^2[1 - M_2(M_2 - c_2)] = 1.82$.[8] Mean squared errors resulting from this truncated Gaussian prior are reported in the dotted lines in Figure 8. These uniformly dominate those for the simple $N(0,5)$ prior.

To summarize, the $p > 0.5$ normalization introduces substantial distortions in the Bayesian posterior distribution that can be largely avoided with other normalizations. These distortions may turn out favorably for purposes of generating a point estimate of p itself, so that if p is the only parameter of interest, the normalization might be desired on these grounds. Notwithstanding, confidence regions for p that result from this approach are not to be trusted. By contrast, normalization based on the identification principle seems to produce substantially superior point estimates for the other parameters and much better coverage probabilities in almost all cases. Moreover, one should check to make sure that the prior used is sensible given the normalization that is to be adopted– what functions as a

[8] See for example Maddala (1983, pp. 365-366).

vague prior for one normalization can be significantly distorting with another normalization.

4 Structural VAR's.

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

$$\mathbf{B}_0\mathbf{y}_t = \mathbf{k} + \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \cdots + \mathbf{B}_p\mathbf{y}_{t-p} + \mathbf{u}_t \qquad (8)$$

where $\mathbf{u}_t \sim N(\mathbf{0}, \mathbf{D}^2)$ with $\mathbf{D}$ a diagonal matrix. A structural VAR typically makes both exclusion restrictions and normalization conditions on $\mathbf{B}_0$ in order to be identified. To use a familiar example (e.g., Hamilton, 1994, pages 330-331), let q_t denote 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

$$q_t = \beta p_t + \boldsymbol{\delta}_1'\mathbf{x}_t + u_{1t} \qquad (9)$$

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

$$q_t = \gamma p_t + hw_t + \boldsymbol{\delta}_2^{\prime}\mathbf{x}_t + u_{2t},$$

with the supply elasticity expected to be positive ($\gamma > 0$) and freezing weather should discourage orange production ($h < 0$). We might also use an equation for weather of the form $w_t = \boldsymbol{\delta}_3^{\prime}\mathbf{x}_t + u_{3t}$, where perhaps $\boldsymbol{\delta}_3 = \mathbf{0}$. This system is an example of (8) 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):

$$\mathbf{B}_0 = \begin{bmatrix} 1 & -\beta & 0 \\ 1 & -\gamma & -h \\ 0 & 0 & 1 \end{bmatrix}. \tag{10}$$

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 is 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:

$$\begin{bmatrix} 1 & 2 & 0 \\ 1 & -0.5 & 0.5 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \begin{bmatrix} 0.8 & 1.6 & 0 \\ 1.2 & -0.6 & 0.6 \\ 0 & 0 & 1.8 \end{bmatrix} \begin{bmatrix} q_{t-1} \\ p_{t-1} \\ w_{t-1} \end{bmatrix} +$$

$$\begin{bmatrix} 0 & 0 & 0 \\ -0.8 & 0.4 & -0.4 \\ 0 & 0 & -0.9 \end{bmatrix} \begin{bmatrix} q_{t-2} \\ p_{t-2} \\ w_{t-2} \end{bmatrix} + \begin{bmatrix} u_{dt} \\ u_{st} \\ u_{wt} \end{bmatrix}. \tag{11}$$

In this example, the true demand elasticity $\beta = -2$ and supply elasticity $\gamma = 0.5$, while $\mathbf{D} = \mathbf{I}_3$. Demand shocks are AR(1) with exponential decay factor 0.8 while supply and weather shocks are AR(2) with damped sinusoidal decay.

Figure 9 shows contours of the concentrated log likelihood for a sample of size $T = 50$ from this system.[9] 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 $\beta = -2.09$, $\gamma = 0.28$, and $h = -0.69$, close to the true values, and corresponding to the hill in the upper left triangle of the bottom left panel in Figure 9. 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 ($\beta = -2, \gamma = 0.5$) and a second smaller mass around their flipped values ($\beta = 0.5, \gamma = -2$). Although the posterior probabilities associated with the former are much larger than the latter, the likelihood function merges 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 parameters under this normalization.

To describe this distribution systematically, we generated 1000 samples $\{\mathbf{y}_t\}_{t=1}^{T}$ each

[9] The likelihood has been concentrated by first regressing q_t and p_t on y_{t-1} and y_{t-2}, and regressing w_t on w_{t-1} and w_{t-2}, to get a residual vector $\hat{\mathsf{u}}_t$ and then evaluating at the true $\mathsf{D} = \mathsf{I}_3$. That is, for $\mathsf{B}_0(\beta, \gamma, h)$ the matrix in (10), we evaluated

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

of size $T = 50$ from this model, and generated 100 draws from the posterior distribution of $(\beta, \gamma, h, d_1, d_2, d_3 | \mathbf{y}_1, ..., \mathbf{y}_T)$ for each sample using a diffuse prior; see Appendix B for details on the algorithm used to generate these draws. This is analogous to what an applied researcher would do in order to calculate standard errors if the maximum likelihood estimates for the researcher's single observed sample happened to equal exactly the true values that had actually generated the data. 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 upper left panel of Figure 10. It is instructive (though not standard practice) to examine the actual probability distribution underlying this familiar plot. The upper left panel of Figure 11 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-unit shock to demand is interpreted as something that raises the price by 0.5, though in a significant minority of the draws, a one-unit shock to demand is interpreted as something that lowers the price by 0.5. 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 top right panel of Figure 11). 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 highlighted in Figure 9 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 (10) becomes noninvertible 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 (8) by

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

We then have

$$\begin{bmatrix} 1 & 0 & \pi_1 \\ 0 & 1 & \pi_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \mathbf{\Pi}_1 \mathbf{y}_{t-1} + \mathbf{\Pi}_2 \mathbf{y}_{t-2} + \begin{bmatrix} v_{1t} \\ v_{2t} \\ v_{3t} \end{bmatrix}. \tag{12}$$

Figure 12 plots likelihood contours for this parameterization as a function of π_1, π_2, and ρ, the correlation between v_{1t} and v_{2t}.[10] Although this is exactly the same sample of data displayed in Figure 9, the likelihood function for this parameterization is perfectly well behaved, with a unique mode near the population values of $\pi_1 = 0.4$, $\pi_2 = -0.2$, and $\rho = -0.51$. Indeed, (12) 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

[10] We set $E(v_{1t}^2) = 0.68$ and $E(v_{2t}^2) = 0.32$, their population values.

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 simulations described above are [0.00,0.71] for π_1 and [-0.42,0.04] for π_2.

Although this π-normalization eliminates the egregious problems associated with the β-normalization in (10), 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 (12) by

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

to obtain

$$\begin{bmatrix} 1 & -\beta & \pi_1 - \beta\pi_2 \\ 1 & -\gamma & \pi_1 - \gamma\pi_2 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} q_t \\ p_t \\ w_t \end{bmatrix} = \mathbf{B}_1\mathbf{y}_{t-1} + \mathbf{B}_2\mathbf{y}_{t-2} + \begin{bmatrix} v_{1t} - \beta v_{2t} \\ v_{1t} - \gamma v_{2t} \\ v_{3t} \end{bmatrix}. \qquad (13)$$

Comparing (13) with (11), the structural parameter β must be chosen so as to make the (1,3) element of $\mathbf{B}_0$ zero, or

$$\beta = \pi_1/\pi_2. \qquad (14)$$

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 (14). 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 (see Figure 13). 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 (10) as

$$\mathbf{B}_0 = \begin{bmatrix} -\eta & 1 & 0 \\ 1 & -\gamma & -h \\ 0 & 0 & 1 \end{bmatrix} \tag{15}$$

and concluding that η may be near zero.

When we performed the analogous 100,000 simulations for the η-normalization (15), the 95% confidence interval for η is [-1.88,0.45], a more convenient and accurate way to summarize the basic fact that the demand curve is relatively steep, with elasticity $\beta = \eta^{-1} > -0.53$ and possibly even vertical or positively sloped. The response of price to a demand shock for this normalization is plotted in the upper-right panel of Figure 10. The bimodality of the distribution of $\psi_{12}(0)$ and enormous tails of $\psi_{12}(1)$ have both disappeared (second row of Figure 11).

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 upper right panel of Figure 10, multiplying each value of $\partial y_{2,t+k}/\partial u_{1t}^{*}$ by the constant -0.5 before plotting the figure so as to get a value that corresponds to the identical concept and scale as the one measured in the upper left panel of Figure 10. The difference between this harmless rescaling (multiplying by the constant -0.5) and the issue of normalization discussed in this paper is that the upper-left panel of Figure 10 is the result of multiplying $\partial y_{2,t+k}/\partial u_{1t}^{*}$ not by the constant -0.5 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 first panel of Figure 10, all over the map.

Although the η-normalization would seem to offer a better way to summarize what the data have to say about the slope of the demand curve and effects of shocks to it, it does nothing about the fragility of the estimate of γ. Moreover, the particular approach followed here of swapping β with η may be harder to recognize or generalize in more complicated examples. It is thus of interest to see how our two automatic solutions for the normalization problem work for this particular example.

To discuss normalization more generally for a structural VAR, we premultiply (8) by $\mathbf{D}^{-1}$ and transpose,

$$\mathbf{y}_t'\mathbf{A}_0 = \mathbf{c} + \mathbf{y}_{t-1}'\mathbf{A}_1 + \mathbf{y}_{t-2}'\mathbf{A}_2 + \ldots + \mathbf{y}_{t-p}'\mathbf{A}_p + \boldsymbol{\varepsilon}_t' \tag{16}$$

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 $E(\boldsymbol{\varepsilon}_t\boldsymbol{\varepsilon}_t') = \mathbf{I}_n$. Identification is typically achieved by imposing zeros on $\mathbf{A}_0$. For the supply-demand example (11),

$$\mathbf{A}_0 = \begin{bmatrix} a_{11} & a_{12} & 0 \\ a_{21} & a_{22} & 0 \\ 0 & a_{32} & a_{33} \end{bmatrix}.$$

The normalization problem arises because, even though the model is identified in the conventional sense from these zero restrictions, multiplying any column of $\mathbf{A}_j$ by -1 for $j = 0, 1, ..., p$ results in the identical value for the likelihood function. For $n = 3$ as here, there are 8 different values of $\mathbf{A}_0$ that work equally well. In any Bayesian or classical analysis that produces a particular draw for $\mathbf{A}_0$ (for example, a single draw from the posterior distribution), we have to choose which of these 8 possibilities to use in constructing our simulation of the range of possible values for $\mathbf{A}_0$.

Let $\boldsymbol{\theta}$ denote the unknown elements of $\mathbf{A}_0$; for this example, $\boldsymbol{\theta} = (a_{11}, a_{12}, a_{21}, a_{22}, a_{32}, a_{33})'$. Waggoner and Zha (2003a) suggested that each simulated draw for $\boldsymbol{\theta}$ should be chosen such that the concentrated log likelihood $\mathcal{L}(\boldsymbol{\theta})$ is finite for all $\boldsymbol{\theta} = \lambda\hat{\boldsymbol{\theta}} + (1 - \lambda)\boldsymbol{\theta}$ for $\hat{\boldsymbol{\theta}}$ the MLE and for all $\lambda \in [0, 1]$. This is implemented as follows. Let $\hat{\mathbf{a}}_k$ denote the kth column of $\mathbf{A}_0(\hat{\boldsymbol{\theta}})$. Let $\mathbf{A}_0$ denote a proposed candidate value for the matrix of contemporaneous coefficients, drawn from a simulation of the Bayesian posterior distribution. The Waggoner-Zha algorithm for deciding whether the candidate $\mathbf{A}_0$ satisfies the normalization condition is to check the sign of $\mathbf{e}_k'\mathbf{A}_0^{-1}\hat{\mathbf{a}}_k$, where $\mathbf{e}_k$ denotes the kth column of $\mathbf{I}_n$. If $\mathbf{e}_k'\mathbf{A}_0^{-1}\hat{\mathbf{a}}_k > 0$, the kth column of $\mathbf{A}_0$ is determined to be correctly normalized. If $\mathbf{e}_k'\mathbf{A}_0^{-1}\hat{\mathbf{a}}_k < 0$, the kth column of $\mathbf{A}_0$ is multiplied by -1.

When using this algorithm to calculate standard errors in a particular application, one has a single observed sample and particular maximum likelihood estimate $\hat{\boldsymbol{\theta}}$ from which the normalization is to be determined. In attempting to evaluate the promise of this method with a broader Monte Carlo investigation as here, in principle one has to take into account the possible sampling distribution of $\hat{\boldsymbol{\theta}}$ itself, though we have found that it does not seem to make much difference how one handles this question in practice. As one way to design the Monte Carlo study, we generated ten different samples, each of size $T = 50$, and found the MLE for each sample. Of course, there are eight equivalent MLE's for each of these 10 samples (corresponding to whether each of the three columns of $\hat{\mathbf{A}}_0$ is multiplied by ± 1), and for each sample we chose as its "MLE" the one of these eight for which $\parallel \hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \parallel$ was smallest, where $\boldsymbol{\theta}_0$ denotes the true values, i.e., the numbers from the left-most matrix in (11). For each of these ten samples, we generated 100 different samples, and for each of these 100 samples used the Waggoner-Zha normalization based on the root sample's $\hat{\mathbf{A}}_0$. For each sample, we calculated 100 draws from the posterior distribution, for a total of 100,000 parameter draws coming from 1000 different samples each of size $T = 50$.

The lower right panel of Figure 10 gives the impulse-response function $\psi_{12}(k)$ for this Waggoner-Zha normalization along with 90% confidence intervals, while the third row of

Figure 11 plots the densities of the first two terms in this function. The results are basically indistinguishable from those for the η-normalization.

The second automatic procedure we investigated is the Wald normalization. For the jth parameter draw $\boldsymbol{\theta}^{(j)}$ as above ($j = 1, ..., 100,000$), we calculated the 8 possible permutations $\left\{\boldsymbol{\theta}^{(j,m)}\right\}_{m=1}^{8}$, and calculated the 8 corresponding Wald test statistics,

$$W_{jm} = \left(\boldsymbol{\theta}^{(j,m)} - \hat{\boldsymbol{\theta}}^{i(j)}\right)' \hat{\mathbf{H}} \left(\boldsymbol{\theta}^{(j,m)} - \hat{\boldsymbol{\theta}}^{i(j)}\right)$$

for $\hat{\mathbf{H}}$ the matrix in (5) and $\hat{\boldsymbol{\theta}}^{i(j)}$ the MLE from the root sample $i \in \{1, ..., 10\}$ associated with draw j. The value $\boldsymbol{\theta}^{(j)}$ was then taken to be the element of the set $\left\{\boldsymbol{\theta}^{(j,m)}\right\}_{m=1}^{8}$ for which W_{jm} is smallest.

The lower left panel of Figure 10 gives the impulse-response function $\psi_{12}(k)$ for this Wald normalization along with 90% confidence intervals, while the fourth row of Figure 11 plots the densities of the first two terms in this function. The results are again indistinguishable from those for either the η-normalization or the Waggoner-Zha normalization.

To be sure, the $\psi_{12}(k)$ function is not estimated all that accurately in this example, even for our preferred normalizations. This is a basic limitation of the data and model– the identification here, though valid, is relatively weak. However, there is no reason to compound the unavoidable problem of weak identification with a normalization that imposes a pathological topography to the likelihood surface, as manifest in Figures 9 and 10(a). Indeed, a suitable normalization is particularly imperative in such cases in order to come away with a clear understanding of which features of the model the data are informative about.

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, and $h < n$ is the number of cointegrating relations 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 nonsingular $(h \times h)$ matrix. Two popular normalizations are to set the first h rows and 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:

$$\begin{aligned} \Delta y_{1t} &= \varepsilon_{1t} \\ \Delta y_{2t} &= y_{1,t-1} - y_{2,t-1} + \varepsilon_{2t} \end{aligned} \qquad (17)$$

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 14 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 much more severe. To illustrate this, we generated $N = 10,000$ samples each of size $T = 50$ from this model with $b_2 = 0.1, a_1 = 1$, and $a_2 = -1$, choosing the values of a_2 and b_2 for each sample so as to maximize the likelihood, given $a_1 = 1$. Figure 15 plots kernel estimates of the small-sample distribution of the maximum likelihood estimates $\hat{a}_2$ and $\hat{b}_2$. 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 Figure 16), with MSE's around 0.8.[11]

One can understand why the normalization that satisfies the identification principle ($b_2 = 0.1$) results in much better behaved estimates for this example by examining the reduced form of the model:

$$\begin{aligned} \Delta y_{1t} &= \varepsilon_{1t} \\ \Delta y_{2t} &= \pi_1 y_{1,t-1} + \pi_2 y_{2,t-1} + \varepsilon_{2t}. \end{aligned} \tag{18}$$

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.

[11] 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.

Consider next the case where the value of b_1 is also unknown:

$$\begin{bmatrix} \Delta y_{1t} \\ \\ \Delta y_{2t} \end{bmatrix} = \begin{bmatrix} b_1 \\ \\ b_2 \end{bmatrix} \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} y_{1,t-1} \\ \\ y_{2,t-1} \end{bmatrix} + \begin{bmatrix} \varepsilon_{1t} \\ \\ \varepsilon_{2t} \end{bmatrix}. \tag{19}$$

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,

$$\Delta \mathbf{y}_t = \mathbf{\Pi} \mathbf{y}_{t-1} + \boldsymbol{\varepsilon}_t \tag{20}$$

where cointegration restricts $\mathbf{\Pi}$ to have unit rank. 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 $b_1 = 0, b_2 = 0.1, a_1 = 1, a_2 = -1$ 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 Figure 17. 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

$$\hat{\mathbf{\Pi}} = \begin{bmatrix} \underset{(0.079)}{0.049} & \underset{(0.078)}{-0.0521} \\ \underset{(0.078)}{0.140} & \underset{(0.085)}{-0.147} \end{bmatrix}.$$

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

$$\underset{(0.078)}{0.140}y_{1t} - \underset{(0.085)}{0.147}y_{2t} \sim \mathrm{I}(0) \tag{21}$$

or that the cointegrating vector is $(1, -1.05)'$. Although $(0.140, -0.147)'$ and $(1, -1.05)'$ represent the identical cointegrating vector, the former is measured in units that have an objective definition, namely, 0.140 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 0.140/0.078 is a test of the null hypothesis that this forecast would not change at all.[12] 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 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.

Any hypothesis about the cointegrating vector can be translated into a hypothesis about $\mathbf{\Pi}$, the latter having the advantage that the small-sample distribution of $\hat{\mathbf{\Pi}}$ is much better behaved than are the distributions of transformations of $\hat{\mathbf{\Pi}}$ that are used in other normalizations. For example, in an n-variable system, one would test the null hypothesis

[12] 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.

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 $\hat{\mathbf{\Pi}}$ estimates across simulated samples. One could further use the $\mathbf{\Pi}$-normalization to describe most other magnitudes of interest, such as calculating forecasts $E(\mathbf{y}_{t+j}|\mathbf{y}_t, \mathbf{y}_{t-1}, ..., \mathbf{\Pi})$ 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 described some of the pitfalls that can arise in describing the small-sample distributions of parameters in a wide variety of econometric models where one has imposed a seemingly innocuous normalization. We have called attention in such settings to the loci in the parameter space along which the model is locally unidentified, 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 problem. Asymptotically, 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 in the vicinity of a troublesome locus. The small-sample 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 identification is relatively weak, i.e., when there is significant probability mass near an unidentified locus.

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.

Where this solution is impossible, we offer another practical guideline that provides an approximate way to do the same thing. Given a set of observationally equivalent parameter values $(\boldsymbol{\theta}_1^{(i)}, \boldsymbol{\theta}_2^{(i)}, \ldots, \boldsymbol{\theta}_M^{(i)})$ that are generated from the jth simulated sample, choose the one that would be associated with the smallest Wald statistic for testing $H_0 : \boldsymbol{\theta}_m^{(i)} = \hat{\boldsymbol{\theta}}_{MLE}$.

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 can not hurt to try several different normalizations to make sure that 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 distributions like those in Figure 3 or Figure 11 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

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

$$\pi(p_1, p_2) \propto p_1^{\alpha_1 - 1} p_2^{\alpha_2 - 1}, \tag{22}$$

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

$$\pi(\mu_1, \mu_2) = \varphi\left(\begin{bmatrix} \bar{\mu}_1 \\ \bar{\mu}_2 \end{bmatrix}, \begin{bmatrix} \varsigma_1^2 & 0 \\ 0 & \varsigma_2^2 \end{bmatrix}\right), \tag{23}$$

where $\varphi(\mathbf{x}, \mathbf{\Omega})$ denotes the normal pdf with mean $\mathbf{x}$ and covariance matrix $\mathbf{\Omega}$ and the restrictions $\overline{\mu}_1 = \overline{\mu}_2$ 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 $\pi(\boldsymbol{\theta}|\mathbf{y})$ 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

$$\pi(s_t \mid y_t, \boldsymbol{\theta}) = \frac{\pi(y_t \mid s_t, \boldsymbol{\theta})\, \pi(s_t \mid \boldsymbol{\theta})}{\sum_{s_t=1}^{2} \pi(y_t \mid s_t, \boldsymbol{\theta})\, \pi(s_t \mid \boldsymbol{\theta})}, \tag{24}$$

with

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

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

$$p_1 + p_2 = 1.$$

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 (22) and (23) with the likelihood function leads to

$$\begin{aligned} \pi(p_1, p_2 \mid \mathbf{y}, \mathbf{s}) &= \pi(p_1, p_2 \mid \mathbf{s}) \\ &\propto \pi(\mathbf{s} \mid p_1, p_2)\, \pi(p_1, p_2) \qquad (25) \\ &\propto p_1^{T_1+\alpha_1-1}\, p_2^{T_2+\alpha_2-1}, \end{aligned}$$

$$\begin{aligned} \pi(\mu_1, \mu_2 \mid \mathbf{y}, \mathbf{s}, p_1, p_2) &\propto \pi(\mathbf{y} \mid \mathbf{s}, p_1, p_2, \mu_1, \mu_2)\, \pi(\mu_1, \mu_2) \\ &= \varphi\left(\begin{bmatrix} \tilde{\mu}_1 \\ \\ \tilde{\mu}_2 \end{bmatrix}, \begin{bmatrix} \frac{\varsigma_1^2}{\varsigma_1^2 T_1 + 1} & 0 \\ \\ 0 & \frac{\varsigma_2^2}{\varsigma_2^2 T_2 + 1} \end{bmatrix} \right), \qquad (26) \end{aligned}$$

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

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

The posterior density (25) is of Beta form and (26) is of Gaussian form; thus, sampling from these distributions is straightforward.

Truncated Gaussian prior.

The truncated Gaussian prior used in the text has the form:

$$\pi(\mu_1, \mu_2) = \pi(\mu_1)\pi(\mu_2|\mu_1), \tag{27}$$

where $\pi(\mu_2|\mu_1)$ is given by (7). Replacing the symmetric prior (23) with the truncated prior (27) leads to the following posterior pdf of μ_1 and μ_2:

$$\pi(\mu_1, \mu_2 \mid \mathbf{y}, \mathbf{s}, p_1, p_2) = \frac{1}{\Phi\left(\frac{\mu_1 - \bar{\mu}_2}{\varsigma_2}\right)} \varphi\left(\begin{bmatrix} \tilde{\mu}_1 \\ \tilde{\mu}_2 \end{bmatrix}, \begin{bmatrix} \frac{\varsigma_1^2}{\varsigma_1^2 T_1 + 1} & 0 \\ 0 & \frac{\varsigma_2^2}{\varsigma_2^2 T_2 + 1} \end{bmatrix}\right) \tag{28}$$

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 (28) is not of any standard form. To sample from (28), 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

$$q(\mu^{(j)}, \mu' \mid \mathbf{y}, \mathbf{s}, p_1, p_2) = \varphi\left(\begin{bmatrix} \mu_1^{(j)} \\ \mu_2^{(j)} \end{bmatrix}, c \begin{bmatrix} \frac{\varsigma_1^2}{\varsigma_1^2 T_1 + 1} & 0 \\ 0 & \frac{\varsigma_2^2}{\varsigma_2^2 T_2 + 1} \end{bmatrix}\right), \tag{29}$$

where c is a scaling factor to be adjusted to maintain an optimal acceptance ratio (e.g., between 25% to 40%). Given the previous posterior draw $\mu^{(j)}$, the algorithm sets $\mu^{(j+1)} = \mu'$ with acceptance probability[13]

[13] Note from (29) that $q(\mu, \mu') = q(\mu', \mu)$, allowing us to use the Metropolis as opposed to the Metropolis-Hastings algorithm.

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

otherwise, the algorithm sets $\mu^{(j+1)} = \mu^{(j)}$. [14]

[14] If the random value $\mu_1^* = \mu_1'$ generated from $q(\mu^{(j)}, \mu' \mid \mathsf{y}, \mathsf{s}, p_1, p_2)$ or $\mu_1^* = \mu_1^{(j)}$ results in a numerical underflow when $\Phi\left(\frac{\mu_1^* - \bar{\mu}_2}{\varsigma_2}\right)$ 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.

Appendix B

All simulations were done using the Gibbs sampler for structural VARs described in Waggoner and Zha (2003b). This technique samples from the posterior distribution associated with the specification given by (16). A flat prior was used to obtain draws of $\mathbf{A}_0, \mathbf{A}_1, \cdots \mathbf{A}_p$ 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[15] , 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. 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.

[15] See Sims and Zha 1994 for a discussion of this result.

Appendix C

Maximum likelihood estimation of (20) can be found using Johansen's (1988) procedure, as described in Hamilton (1994, p. 637). Specifically, let $\hat{\mathbf{\Sigma}}_{\mathsf{vv}} = T^{-1}\sum_{t=1}^{T}\mathbf{y}_{t-1}\mathbf{y}_{t-1}^{\prime}$, $\hat{\mathbf{\Sigma}}_{\mathsf{uu}} = T^{-1}\sum_{t=1}^{T}\Delta\mathbf{y}_{t}\Delta\mathbf{y}_{t}^{\prime}$, $\hat{\mathbf{\Sigma}}_{\mathsf{uv}} = T^{-1}\sum_{t=1}^{T}\Delta\mathbf{y}_{t}\mathbf{y}_{t-1}^{\prime}$, and $\hat{\mathbf{P}} = \hat{\mathbf{\Sigma}}_{\mathsf{vv}}^{-1}\hat{\mathbf{\Sigma}}_{\mathsf{uv}}^{\prime}\hat{\mathbf{\Sigma}}_{\mathsf{uu}}^{-1}\hat{\mathbf{\Sigma}}_{\mathsf{uv}}$. Find $\tilde{\mathbf{a}}_1$, the eigenvector of $\hat{\mathbf{P}}$ associated with the biggest eigenvector and construct $\hat{\mathbf{a}}_1 = \tilde{\mathbf{a}}_1/\sqrt{\tilde{\mathbf{a}}_1^{\prime}\hat{\mathbf{\Sigma}}_{\mathsf{vv}}\tilde{\mathbf{a}}_1}$. The MLE is then $\hat{\mathbf{\Pi}} = \hat{\mathbf{\Sigma}}_{\mathsf{uv}}\hat{\mathbf{a}}_1\hat{\mathbf{a}}_1^{\prime}$.

References

Aitkin, Murray, and Donald B. Rubin (1985). “Estimation and Hypothesis Testing in Finite Mixture Models,” *Journal of the Royal Statistical Society Series B,* 47, pp. 67-75.

Alonso-Borrego, C., and Arellano, M. (1999). “Symmetrically Normalized Instrumental-Variable Estimation Using Panel Data, ” *Journal of Business and Economic Statistics,* 17, pp. 36-49.

Celeux, Gilles, Merilee Hurn, and Christian P. Robert (2000). “Computational and Inferential Difficulties with Mixture Posterior Distributions, *Journal of the American Statistical Association,* 95, pp. 957-970.

Chen, An Mei, Haw-minn Lu, and Robert Hecht-Nielsen (1993). “On the Geometry of Feedforward Neural Network Error Surfaces,” *Neural Computation*, 5(6), pp.910-927.

Chib, Siddhartha, and Edward Greenberg (1996). “Markov Chain Monte Carlo Simulation Methods in Econometrics,” *Econometric Theory,* 12, pp. 409-431.

Frühwirth-Schnatter, Sylvia (2001). “Markov Chain Monte Carlo Estimation of Classical and Dynamic Switching and Mixture Models,” *Journal of the American Statistical Association* 96, pp. 194-209.

Geweke, John (1996). “Bayesian Reduced Rank Regression in Econometrics,” *Journal of Econometrics*, 75, pp. 121-146.

Gregory, A.W., and Veall, M.R. (1985). “Formulating Wald Tests of Nonlinear Restrictions,” *Econometrica*, 53, pp. 465-1468.

Hamilton, James D. (1989). “A New Approach to the Economic Analysis of Nonstationary Time Series and the Business Cycle," *Econometrica,* 57, pp. 357-384.

Hamilton, James D. (1994). *Time Series Analysis.* Princeton, N.J.: Princeton University Press.

Hahn, Jinyong, and Jerry Hausman (2002). "A New Specification Test for the Validity of Instrumental Variables," *Econometrica,* 70, pp. 163-189.

Hauck, Walter W., Jr., and Allan Donner (1977). "Wald's Test as Applied to Hypotheses in Logit Analysis," *Journal of the American Statistical Association,* 72, pp. 851-853.

Johansen, Søren (1988). "Statistical Analysis of Cointegration Vectors," *Journal of Economic Dynamics and Control,* 12, pp. 231-254.

Kleibergen, Frank, and Richard Paap (2002). "Priors, Posteriors, and Bayes Factors for a Bayesian Analysis of Cointegration," *Journal of Econometrics*, 111, pp. 223-249.

Lenk, Peter J., and Wayne S. DeSarbo (2000). "Bayesian Inference for Finite Mixtures of Generalized Linear Models with Random Effects," *Psychometrika* 65, pp.93-119.

Maddala, G.S. (1983). *Limited-Dependent and Qualitative Variables in Econometrics.* Cambridge: Cambridge University Press.

Manski, Charles F. (1988), "Identification of Binary Response Models," *Journal of the American Statistical Association,* 83, pp. 729-738.

Ng, Serena, and Pierre Perron (1997). "Estimation and Inference in Nearly Unbalanced Nearly Cointegrated Systems," *Journal of Econometrics,* 79, pp. 53-81.

Nobile, Agostino, (2000). "Comment: Bayesian Multinomial Problit Models with a Normalization Constraint," *Journal of Econometrics,* 99, pp. 335-345.

Otrok, Christopher, and Charles H. Whiteman (1998). "Bayesian Leading Indicators: Measuring and Predicting Economic Conditions in Iowa," *International Economic Review,* 39, pp. 997-1014.

Pagan, Adrian R., and John Robertson (1997). "GMM and its Problems," Working paper, Australian National University.

Phillips, Peter C. B. (1994). "Some Exact Distribution Theory for Maximum Likelihood Estimators of Cointegrating Coefficients in Error Correction Models," *Econometrica*, 62, pp. 73-93.

Richardson, Sylvia, and Peter J. Green (1997). "On Bayesian Analysis of Mixtures with an Unknown Number of Components," *Journal of the Royal Statistical Society, Series B,* 59, pp. 731-792.

Rothenberg, Thomas J. (1971). "Identification in Parametric Models," *Econometrica*, 39, pp. 577-591.

Ruben, H. (1954). "On the Moments of Order Statistics in Samples from Normal Populations," *Biometrika,* 41, pp. 200-227.

Rüger, Stefan M., and Arnfried Ossen (1996). "Clustering in Weight Space of Feedforward Nets," *Proceedings of the International Conference on Artificial Neural Networks* (ICANN-96, Bochum).

Sims, Christopher A. (1986). "Are Forecasting Models Usable for Policy Analysis?", *Quarterly Review of the Federal Reserve Bank of Minneapolis*, Winter.

Sims, Christopher A., and Tao Zha (1994). "Error Bands for Impulse Responses," Cowles Foundation Discussion Paper No. 1085.

Silverman, B.W. (1992). *Density Estimation for Statistics and Data Analysis.* London: Chapman & Hall.

Smith, Penelope A., and Peter M. Summers (2003). "Identification and Normalization in Markov Switching Models of 'Business Cycles'," Working paper, University of Melbourne.

Stephens, Matthew (2000). "Dealing with Label Switching in Mixture Models," *Journal of the Royal Statistical Society, Series B,* 62, pp. 795-809.

Waggoner, Daniel F., and Tao Zha (2003a). "Likelihood-Preserving Normalization in Multiple Equation Models," *Journal of Econometrics,* 114, pp. 329-347.

Waggoner, Daniel F., and Tao Zha (2003b). "A Gibbs Sampler for Structural Vector Autoregressions," *Journal of Economic Dynamics and Control,* 29, pp. 349-366.

Yogo, Motohiro (2003). "Estimating the Elasticity of Intertemporal Substitution When Instruments Are Weak," forthcoming, *Review of Economics and Statistics.*

Figure Captions

Figure 1. Log likelihood for an i.i.d. $N(0, \sigma^2)$sample of $T = 50$ observations as a function of σ.

Figure 2. Contours of log likelihood for an i.i.d. Gaussian mixture of $T = 50$ observations. True values: $\mu_1 = 1$, $\mu_2 = -1$, $p = 1$.

Figure 3. Posterior densities for i.i.d. mixture normalized according to $p > 0.5$.

Figure 4. Posterior densities for i.i.d. mixture normalized according to $\mu_1 > \mu_2$.

Figure 5. Posterior densities for i.i.d. mixture under Wald normalization.

Figure 6. Average squared difference between posterior mean and true value under three different normalizations as a function of sample size T.

Figure 7. Ninety-percent coverage probabilities under three different normalizations as a function of sample size T.

Figure 8. Mean squared errors under normalization $\mu_1 > \mu_2$ for three different priors as a function of sample size T.

Figure 9. Contours of concentrated log likelihood of structural VAR under the β-normalization.

Figure 10. Impulse-response function and 90% confidence interval for the effect of a one-unit increase in quantity demanded on the price k periods later under four different normalizations.

Figure 11. Posterior densities for the $k = 0$ and $k = 1$ values of the impulse-response functions plotted in Figure 10.

Figure 12. Contours of concentrated log likelihood of structural VAR under the π-normalization.

Figure 13. Effect of freezing weather when the demand curve is steep.

Figure 14. Contours of log likelihood for cointegrated system under three different normalizations.

Figure 15. Sampling densities of maximum likelihood estimates $\hat{a}_2$ and $\hat{b}_2$ for a cointegrated system under the normalization $a_1 = 1$.

Figure 16. Sampling densities of maximum likelihood estimates $\hat{a}_1$ and $\hat{a}_2$ for a cointegrated system under the normalization $b_2 = 0.1$.

Figure 17. Sampling densities of maximum likelihood estimates of elements of the $\mathbf{\Pi}$ matrix in equation (20).

Figure 1

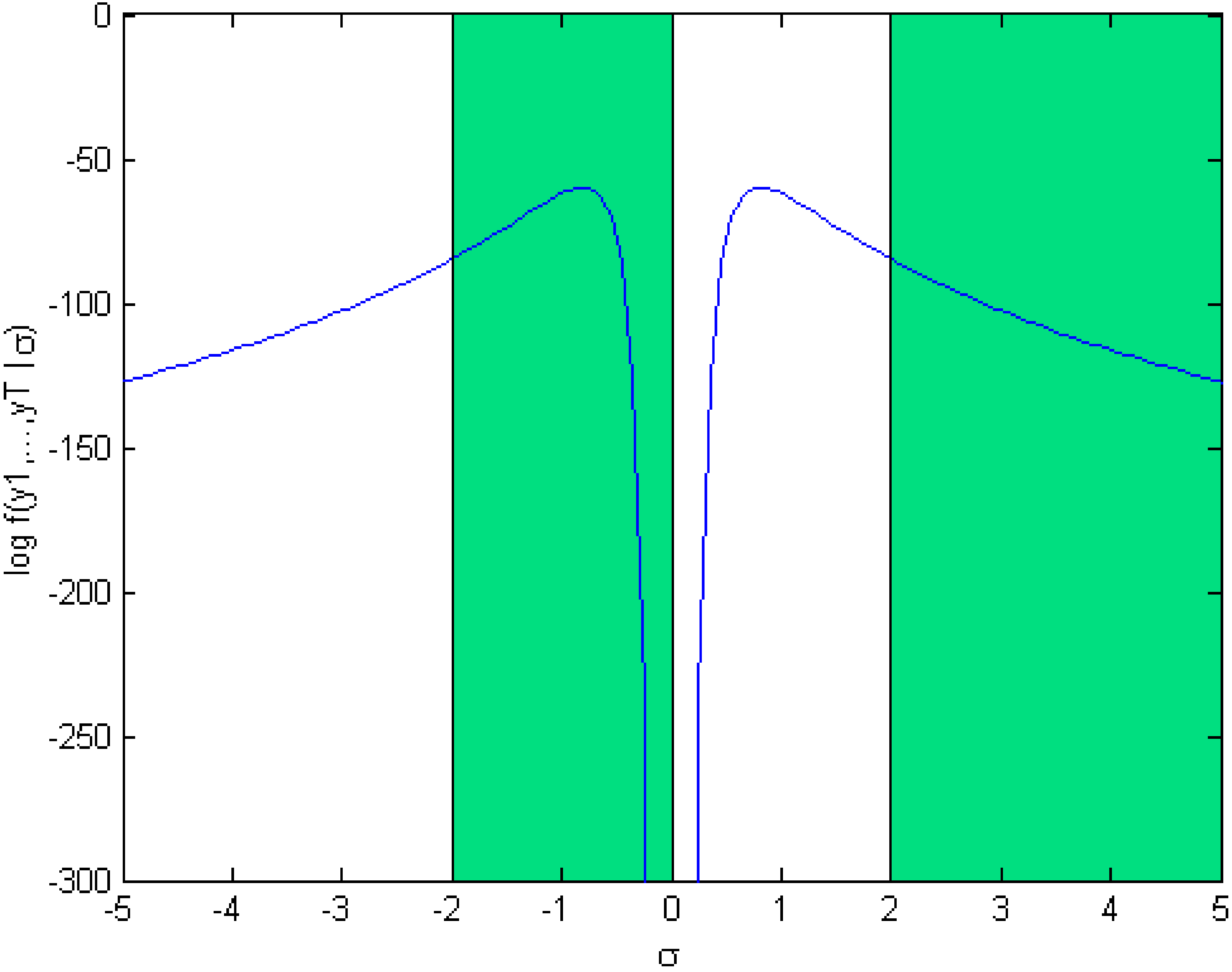


Figure 2

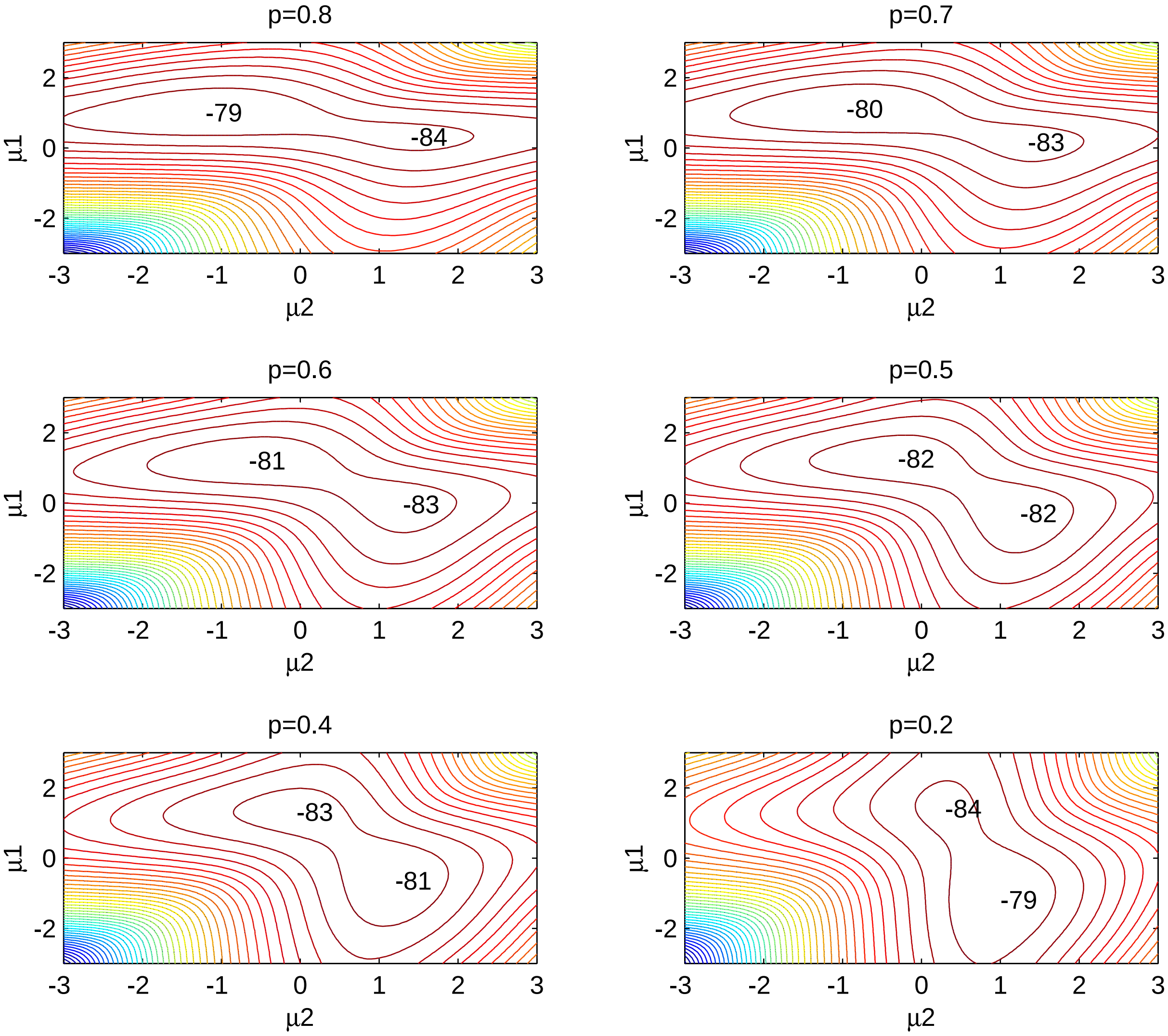
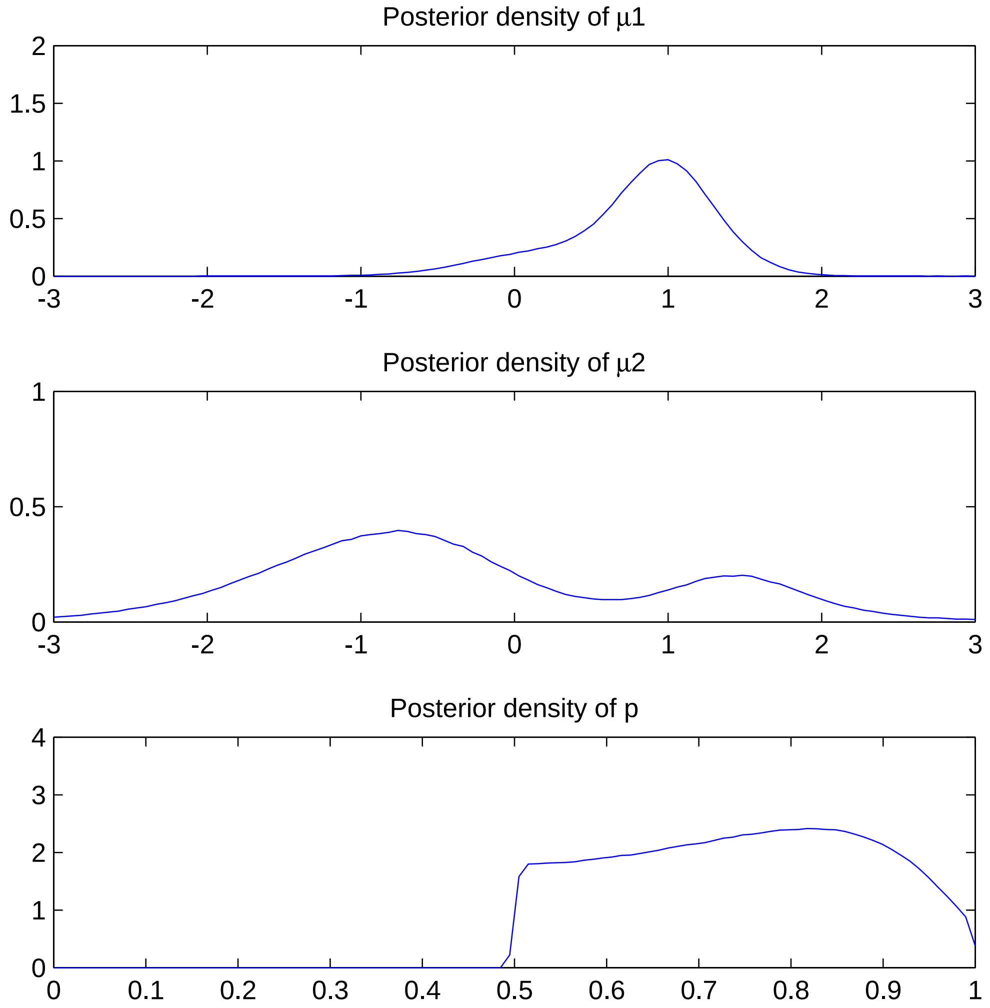

p=0.8
-79
-84
p=0.7
-80
-83
p=0.6
-81
-83
p=0.5
-82
-82
p=0.4
-83
-81
p=0.2
-84
-79
μ1
μ2
2
0
-2
-3
-2
-1
0
1
2
3

Figure 3

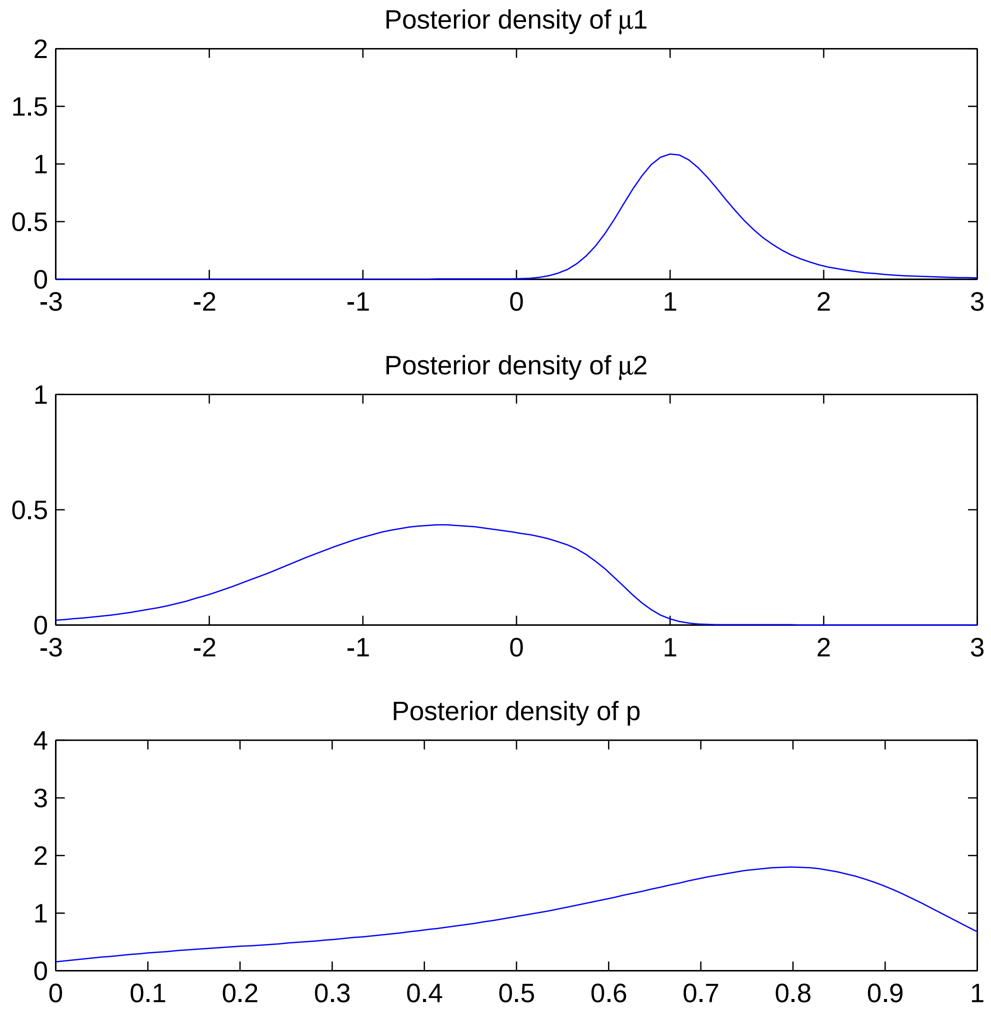


Posterior density of μ1

Posterior density of μ2

Posterior density of p

Figure 4

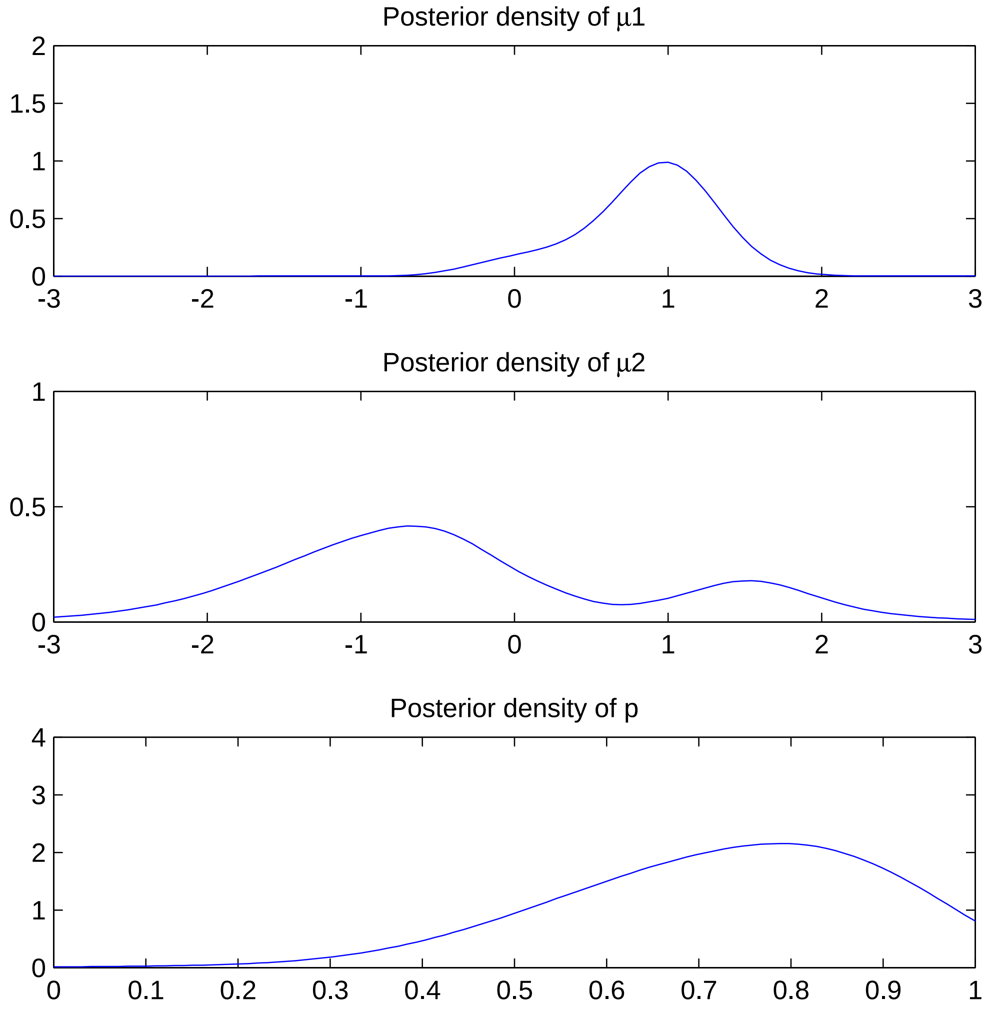


Posterior density of μ1

Posterior density of μ2

Posterior density of p

Figure 5


Posterior density of μ1

Posterior density of μ2

Posterior density of p

Figure 6

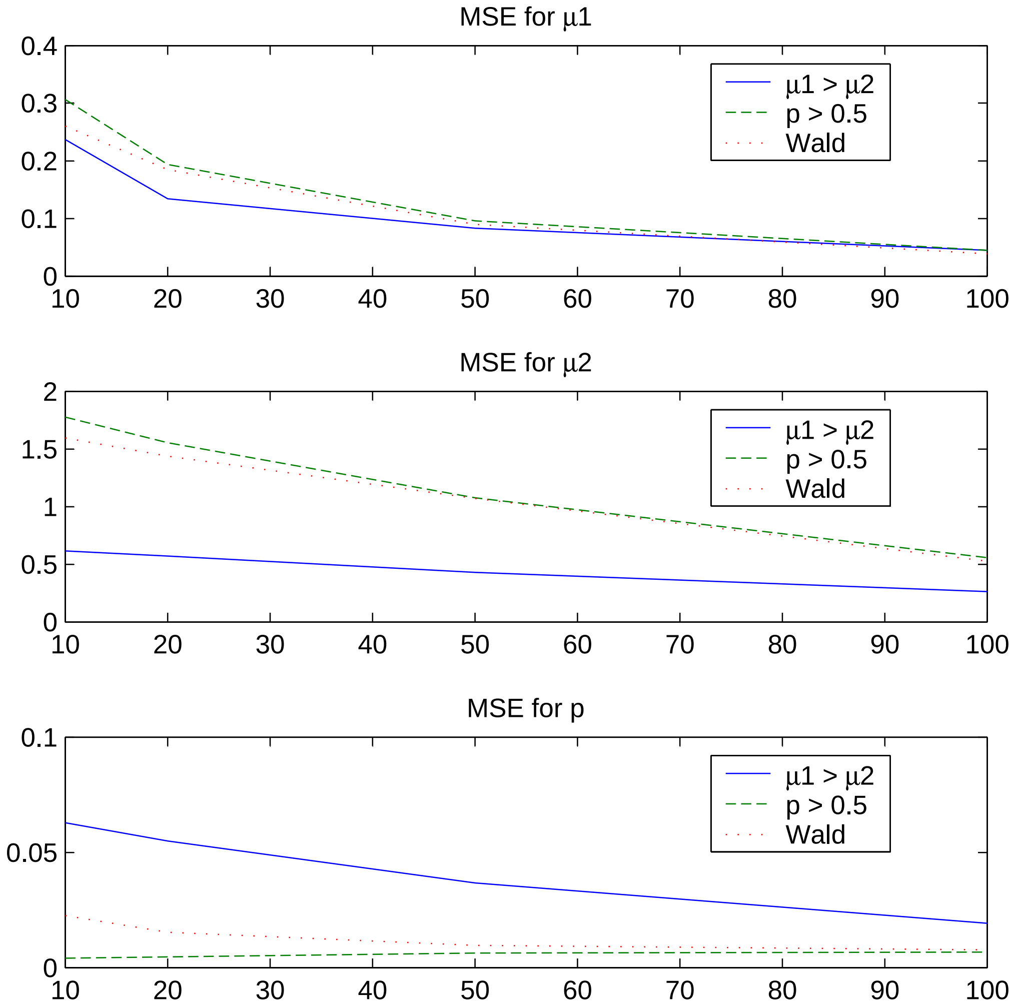
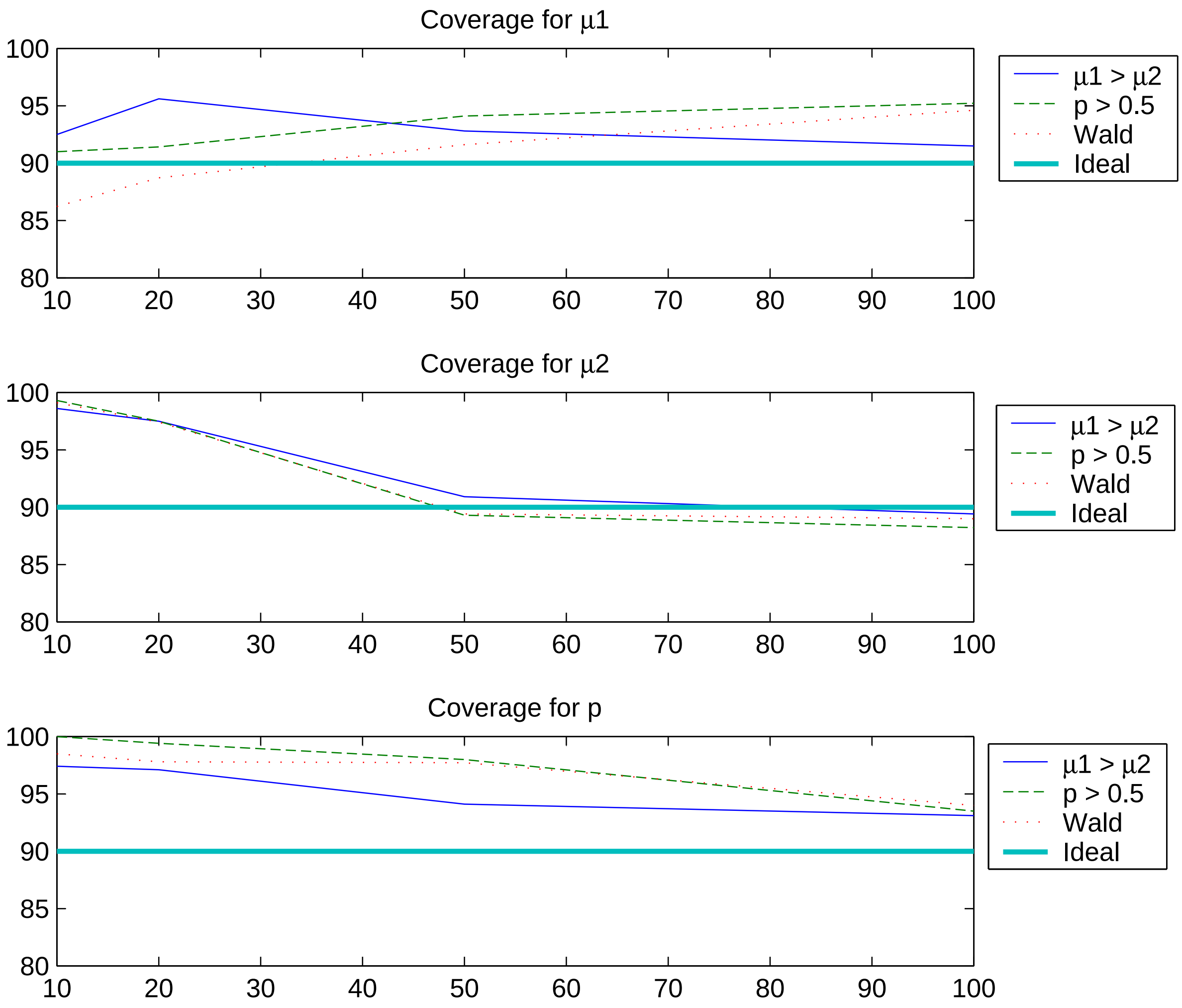

MSE for μ1
0.4
0.3
0.2
0.1
0
10
20
30
40
50
60
70
80
90
100
μ1 > μ2
p > 0.5
Wald
MSE for μ2
2
1.5
1
0.5
0
10
20
30
40
50
60
70
80
90
100
μ1 > μ2
p > 0.5
Wald
MSE for p
0.1
0.05
0
10
20
30
40
50
60
70
80
90
100
μ1 > μ2
p > 0.5
Wald

Figure 7

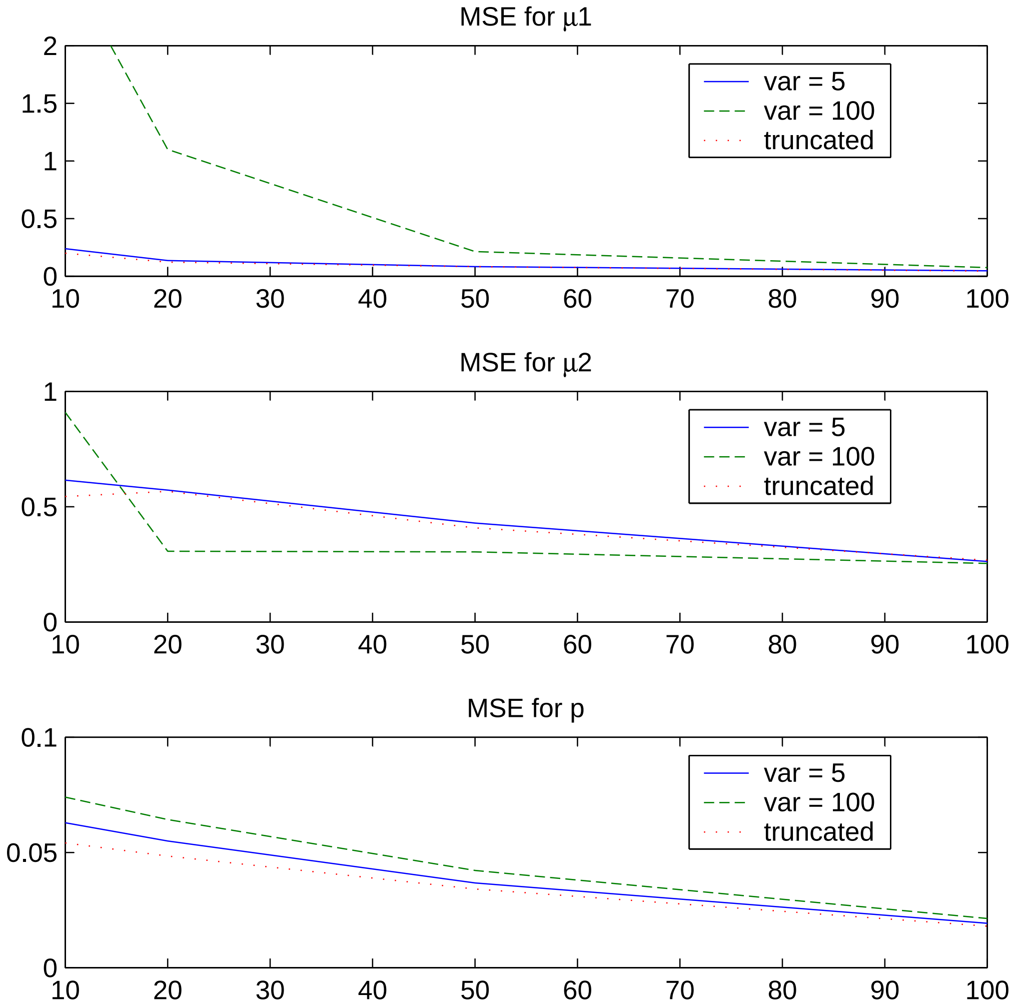


Coverage for µ1

Coverage for µ2

Coverage for p

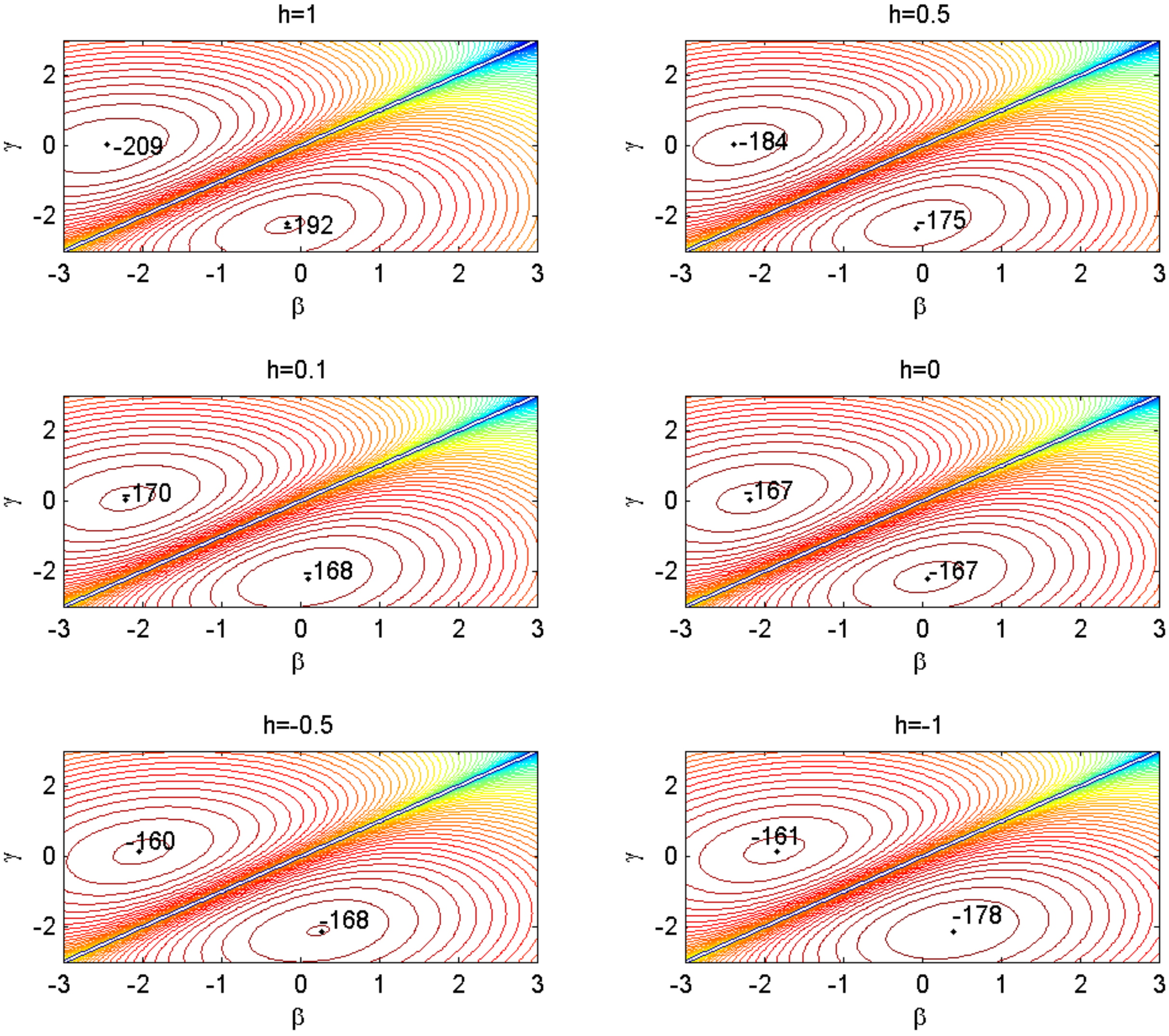
Figure 8


MSE for μ1

MSE for μ2

MSE for p

Figure 9



h=1

h=0.5

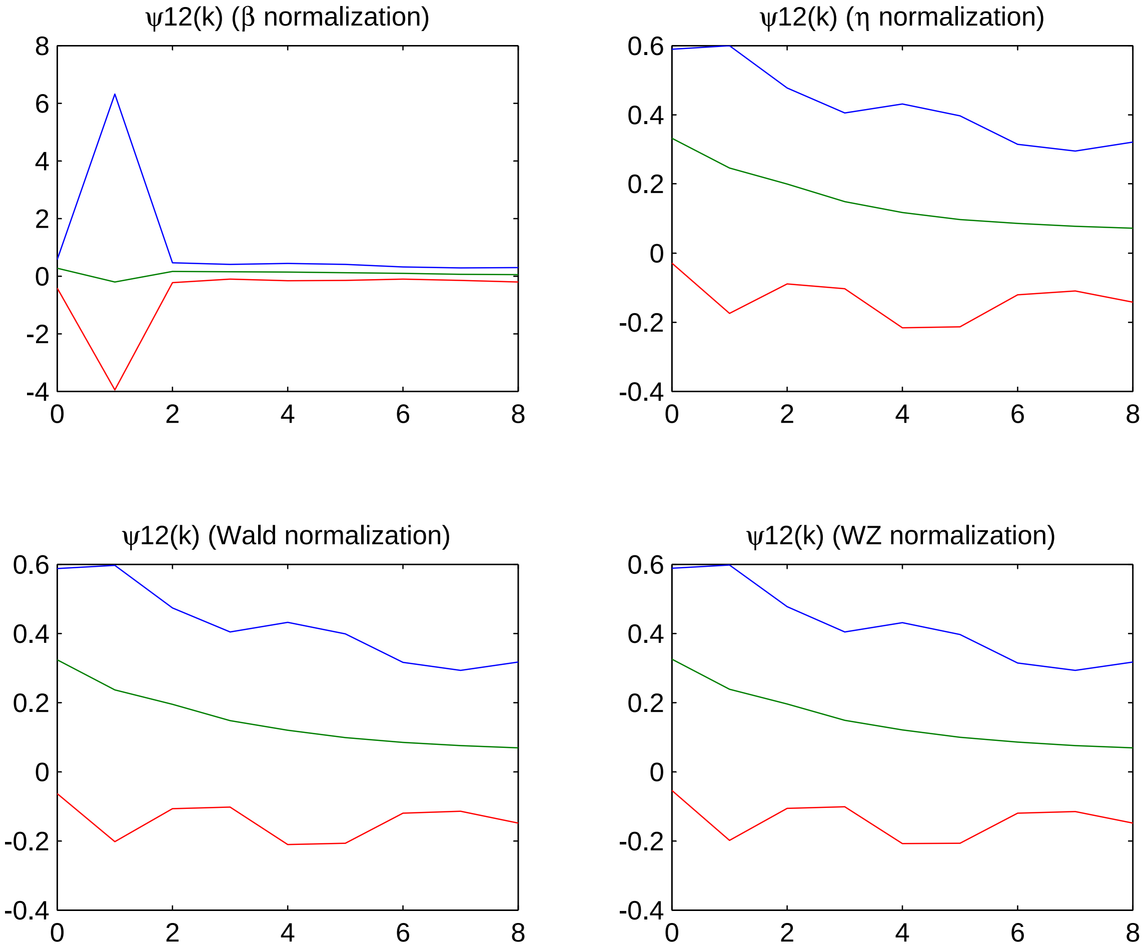
h=0.1

h=0

h=-0.5

h=-1

Figure 10



ψ12(k) (β normalization)

ψ12(k) (η normalization)

ψ12(k) (Wald normalization)

ψ12(k) (WZ normalization)

Figure 11

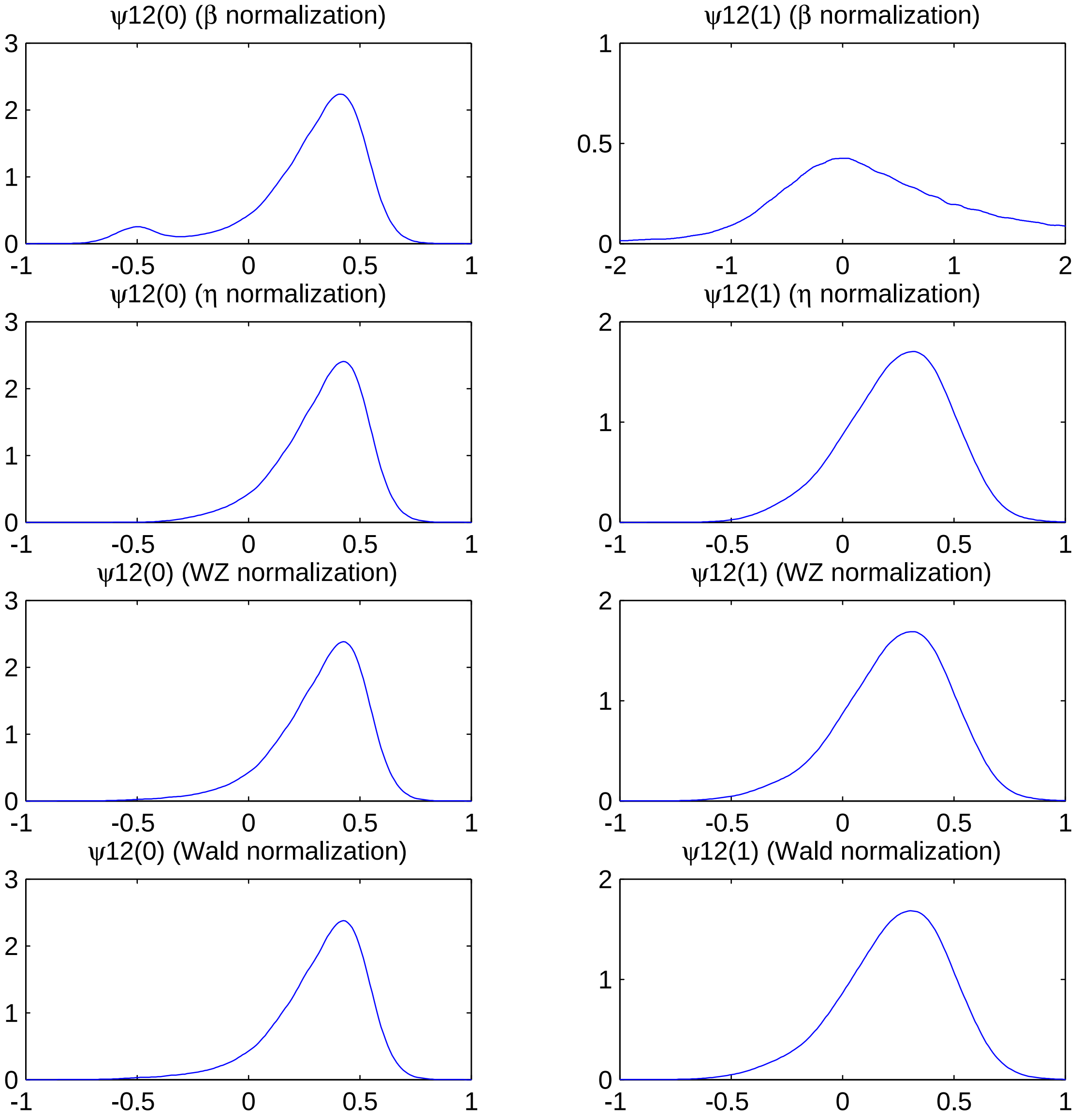
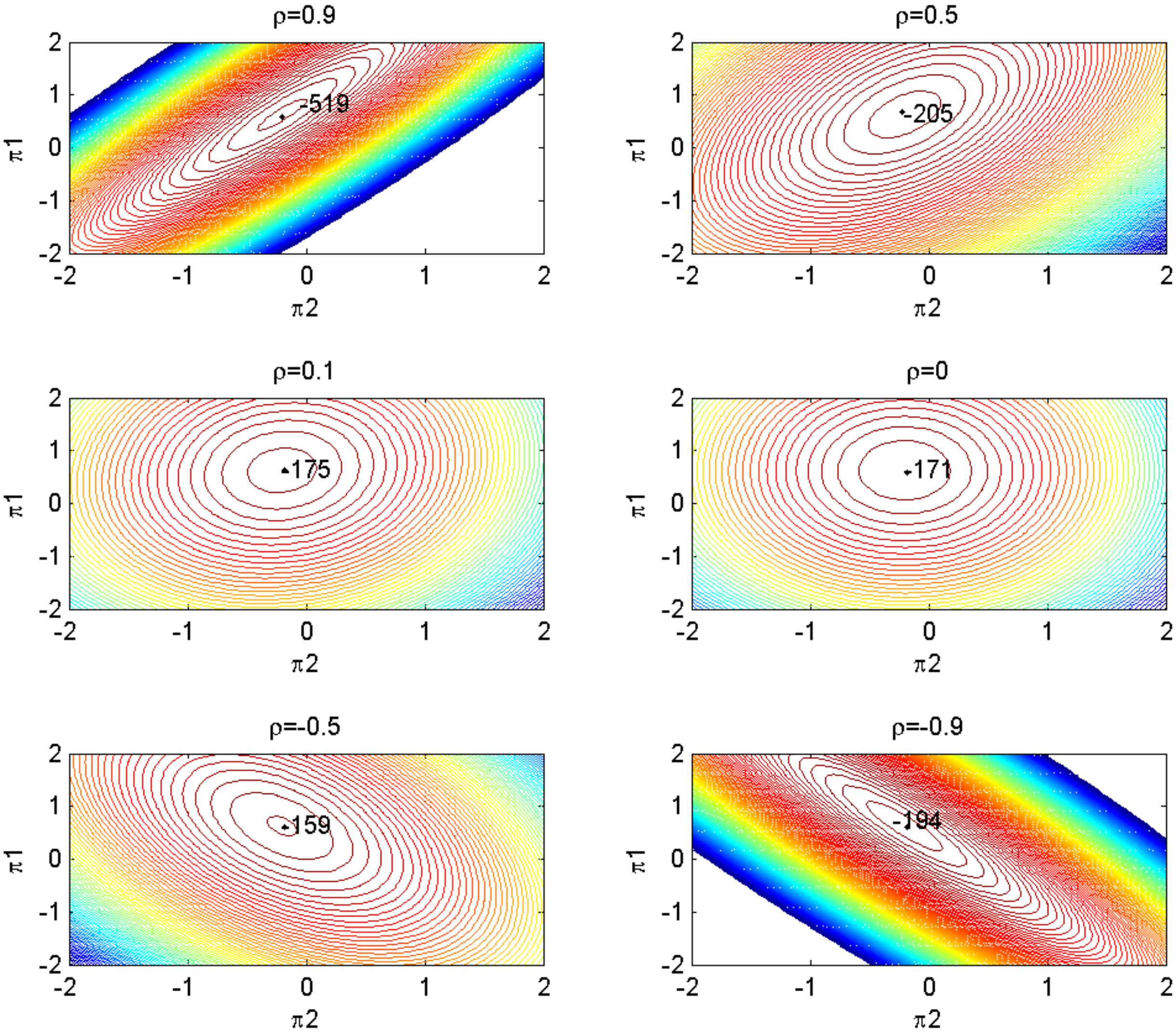

ψ12(0) (β normalization)
3
2
1
0
-1
-0.5
0
0.5
1
ψ12(1) (β normalization)
1
0.5
0
-2
-1
0
1
2
ψ12(0) (η normalization)
3
2
1
0
-1
-0.5
0
0.5
1
ψ12(1) (η normalization)
2
1
0
-1
-0.5
0
0.5
1
ψ12(0) (WZ normalization)
3
2
1
0
-1
-0.5
0
0.5
1
ψ12(1) (WZ normalization)
2
1
0
-1
-0.5
0
0.5
1
ψ12(0) (Wald normalization)
3
2
1
0
-1
-0.5
0
0.5
1
ψ12(1) (Wald normalization)
2
1
0
-1
-0.5
0
0.5
1

Figure 12



ρ=0.9

ρ=0.5

ρ=0.1

ρ=0

ρ=-0.5

ρ=-0.9

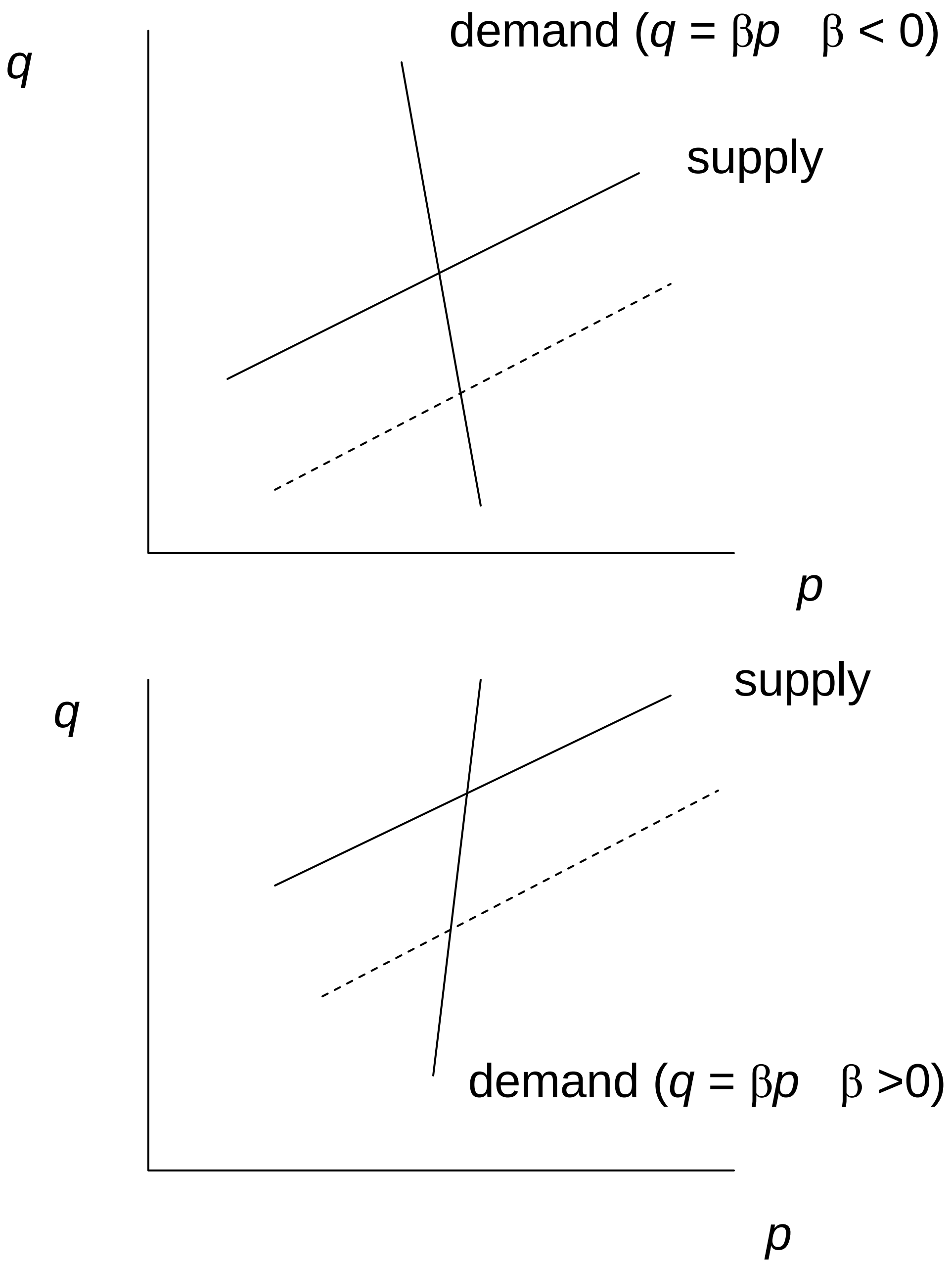

demand ($q = \beta p \quad \beta < 0$)
q
supply
p
supply
q
demand ($q = \beta p \quad \beta > 0$)
p

Figure 13

Figure 14

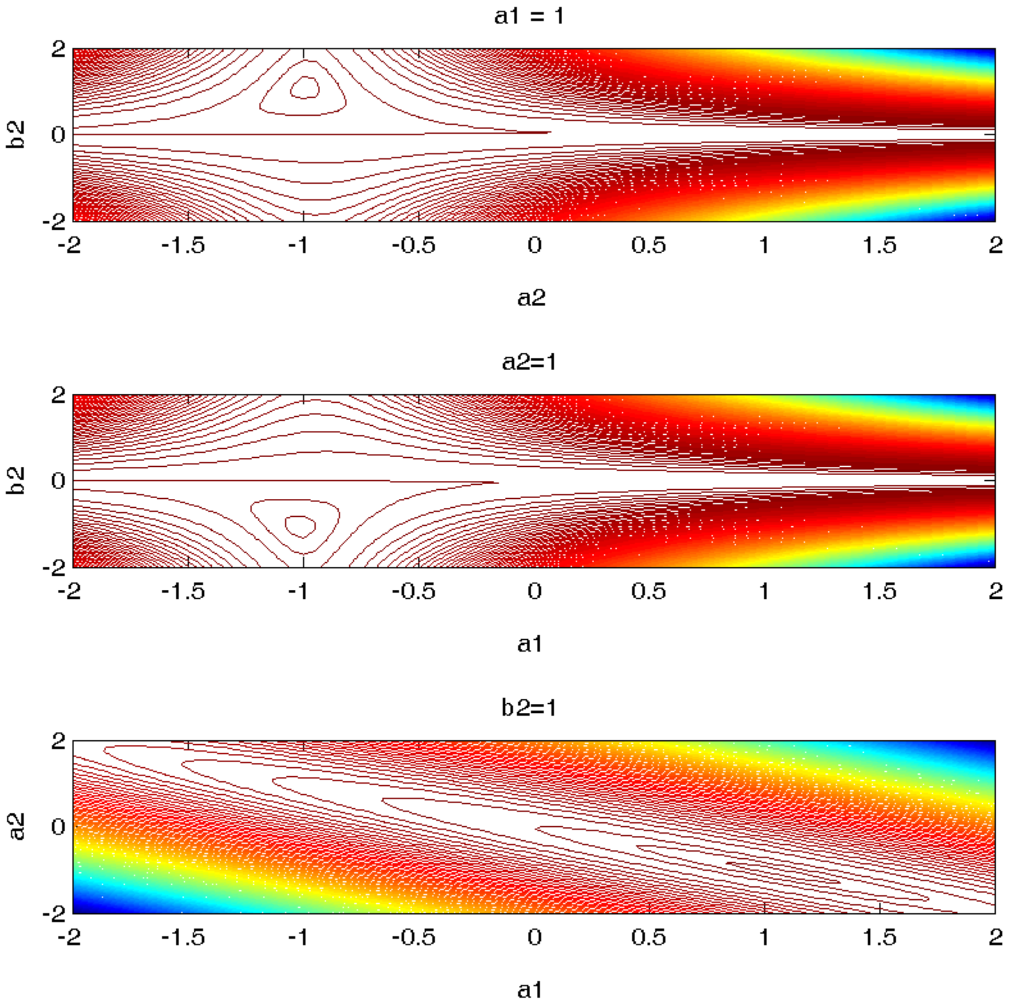
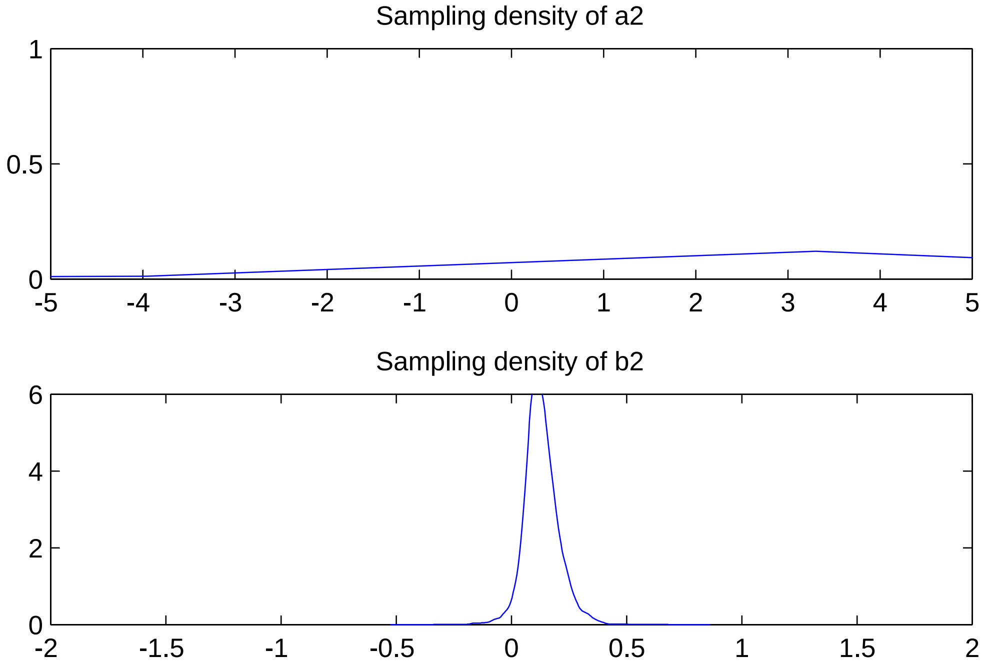

a1 = 1
2
b2 0
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
a2
a2=1
2
b2 0
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
a1
b2=1
2
a2 0
-2
-2 -1.5 -1 -0.5 0 0.5 1 1.5 2
a1

Figure 15



Sampling density of a2

Sampling density of b2

Figure 16

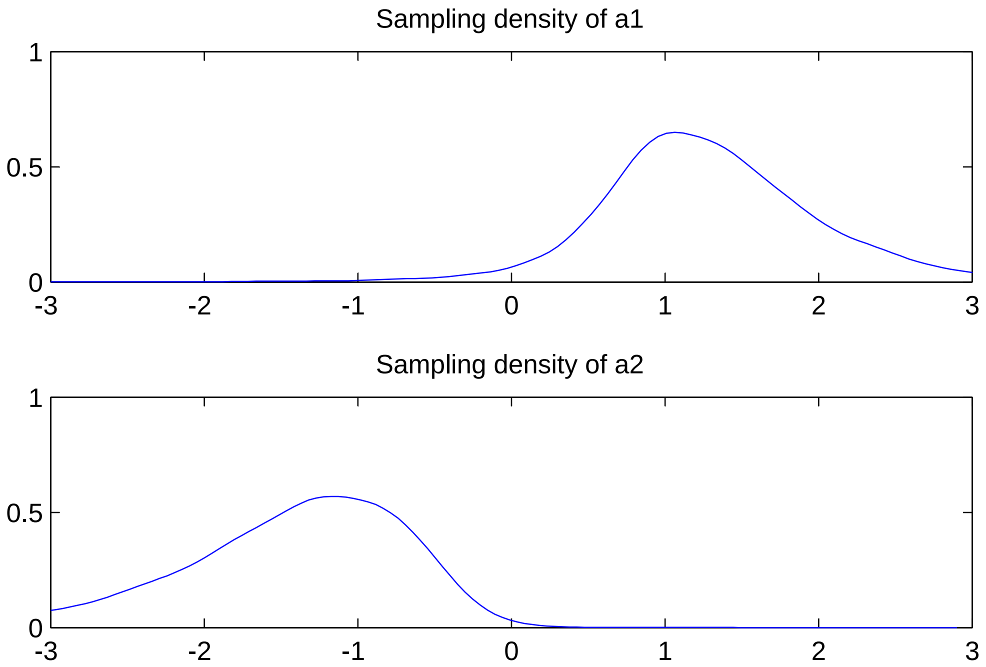
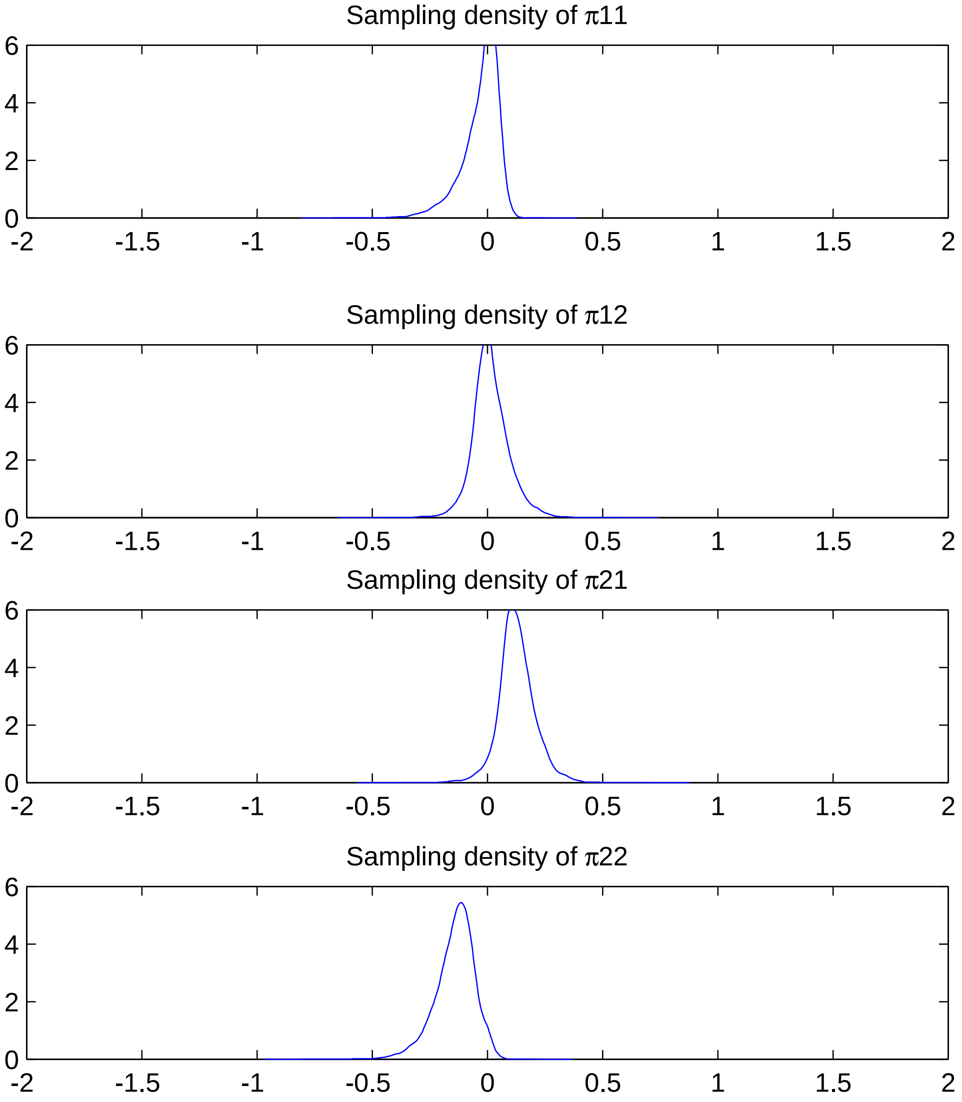

Sampling density of a1
1
0.5
0
-3
-2
-1
0
1
2
3
Sampling density of a2
1
0.5
0
-3
-2
-1
0
1
2
3

Figure 17



Sampling density of π11

Sampling density of π12

Sampling density of π21

Sampling density of π22