Normalization, Probability Distribution, and Impulse Responses

Daniel F. Waggoner and Tao Zha

Federal Reserve Bank of Atlanta
Working Paper 97-11
November 1997

Abstract: When impulse responses in dynamic multivariate models such as identified VARs are given economic interpretations, it is important that reliable statistical inferences be provided. Before probability assessments are provided, however, the model must be normalized. Contrary to the conventional wisdom, this paper argues that normalization, a rule of reversing signs of coefficients in equations in a particular way, could considerably affect the shape of the likelihood and thus probability bands for impulse responses. A new concept called ML distance normalization is introduced to avoid distorting the shape of the likelihood. Moreover, this paper develops a Monte Carlo simulation technique for implementing ML distance normalization.

JEL classification: C32, E52

Key words: ML distance normalization, likelihood, Monte Carlo method, posterior, impulse responses

The authors have benefited from discussions with John Geweke, Lars Hansen, Chuck Whiteman, and especially Chris Sims. The views expressed here are those of 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 Daniel F. Waggoner, Senior Quantitative Analyst, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, 404/521-8278, 404/521-8956 (fax), daniel.f.waggoner@atl.frb.org, or Tao A. Zha, Economist, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, 404/521-8353, 404/521-8956 (fax), tzha@mindspring.com.

Questions regarding subscriptions to the Federal Reserve Bank of Atlanta working paper series should be addressed to the Public Affairs Department, Federal Reserve Bank of Atlanta, 104 Marietta Street, N.W., Atlanta, Georgia 30303-2713, 404/521-8020. The full text of this paper may be downloaded (in PDF format) from the Atlanta Fed's World-Wide Web site at http://www.frbatlanta.org/publica/work_papers/.

Normalization, Probability Distribution, and Impulse Responses

1. Introduction

When dynamic multivariate models such as vector autoregressions (VARs) are used for policy analysis, economically meaningful restrictions are often placed on individual equations or blocks of equations and economic interpretations are usually presented by impulse responses.[1] It is well known that these interpretations are not affected by reversing signs of all coefficients in an equation or equivalently reversing the sign of an identified shock. Consequently, a traditional approach is to restrict arbitrarily the value of one non-zero coefficient to be, say, positive. Textbooks associate this arbitrary approach with formal name “a normalization rule” (e.g., Judge et al., 1985, p.576).

In a recent paper, Sims and Zha (1995) avail themselves of such a normalization rule as though probability bands for impulse responses would also be invariant to how an equation is normalized. Unfortunately, such invariant property no longer holds when one makes probability statements. This paper shows that probability distributions for impulse responses can be sensitive to different normalization rules. It introduces a normalization rule called “ML distance normalization” and argues from a viewpoint of likelihood principles that this rule avoids false inferential conclusions by preserving the shape of the likelihood or the posterior distribution. Moreover, this paper develops a random walk Metropolis algorithm that proves efficient for identified VAR models and applies this algorithm to an example to highlight practical importance of ML distance normalization.

[1] See, for example, Sims (1986), Gordon and Leeper (1994), Leeper, Sims and Zha (1996), and Bernanke, Gertler and Watson (1997).

Section 2 of this paper sets up a general framework of dynamic multivariate models and shows that reversing signs of coefficients in an equation is equivalent to reversing the sign of an identified shock in that equation. Section 3 discusses rigorously the concept of normalization in a new context, introduces the notion of ML distance normalization, and offers analysis of different normalization rules and their implications for statistical inferences. Section 4 develops a Monte Carlo Bayesian method for computing probability bands for impulse responses and applies it to an example to show that other rules of normalization, as compared to ML distance normalization, can yield misleading results in practice. Section 5 closes this paper with concluding remarks.

2. General Framework

This section provides a general framework that summarizes the identified VARs both without any priors and with proper priors. Consider dynamic, stochastic models of the form:[2]

$$A(L)y(t)+C=\varepsilon(t)\ ,\ t=1,\ldots,T\ , \tag{1}$$

where $A(L)$ is an $n\times n$ matrix polynomial in lag operator L, $y(t)$ is an $n\times 1$ vector of observations of n variables at time t, C is an $n\times 1$ vector of constant terms, and $\varepsilon(t)$ is an $n\times 1$ vector of *i.i.d.* innovations so that

$$E\varepsilon(t)=0\,,\quad E\varepsilon(t)\varepsilon(t)'=\underset{n\times n}{\mathbf{I}}\ ,\ \text{all}\ t\ . \tag{2}$$

Following Sims and Zha (1997), rewrite (1) in the matrix form:

$$\mathbf{YA}-\mathbf{XA}_{+}=\mathbf{E}\ , \tag{3}$$

$$\underset{T\times n}{\mathbf{Y}}=\begin{bmatrix} y_1(1) & \cdots & y_n(1) \\ \vdots & \ddots & \vdots \\ y_1(T) & \cdots & y_n(T) \end{bmatrix},\ \underset{T\times n}{\mathbf{E}}=\begin{bmatrix} \varepsilon_1(1) & \cdots & \varepsilon_n(1) \\ \vdots & \ddots & \vdots \\ \varepsilon_1(T) & \cdots & \varepsilon_n(T) \end{bmatrix},$$

[2] Although only constant terms are considered here, the analysis can be extended to other sets of exogenous

$$\underset{T\times k}{\mathbf{X}} = -\begin{bmatrix} y_1(0) & \dots & y_n(0) & \cdots & y_1(1-p) & \cdots & y_n(1-p) & 1 \\ \vdots & \ddots & \vdots & & \vdots & \ddots & \vdots & \vdots \\ y_1(T-1) & \dots & y_n(T-1) & \cdots & y_1(T-p) & \cdots & y_n(T-p) & 1 \end{bmatrix},$$

where p (>0) is a lag length, $k = np+1$, $\mathbf{A}$ is an $n\times n$ matrix, and $\mathbf{A}_+$ is a $k\times n$ matrix. Note that the columns in $\mathbf{A}$ and $\mathbf{A}_+$ correspond to the parameters in individual equations.

Let $\mathbf{a}=(a_i)_{1\le i\le n^2}$ be a vector in $\mathbf{R}^{n^2}$, formed by stacking the columns of $\mathbf{A}$ so that

$$\mathbf{A} \equiv M(\mathbf{a}) = \begin{bmatrix} a_1 & a_{1+n} & a_{1+2n} & \cdots & a_{1+(n-1)n} \\ a_2 & a_{2+n} & a_{2+2n} & \cdots & a_{2+(n-1)n} \\ a_3 & a_{3+n} & a_{3+2n} & \cdots & a_{3+(n-1)n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_n & a_{n+n} & a_{n+2n} & \cdots & a_{n+(n-1)n} \end{bmatrix}.$$

Similarly, let $\mathbf{a}_+$ be a $kn\times 1$ vector formed by stacking the columns of $\mathbf{A}_+$. Note that $M(\cdot)$ is an operation of stacking an $n^2\times 1$ vector $\mathbf{a}$ back to the form of an $n\times n$ matrix. This notation is frequently used in the formal discussion of normalization in the next section.

In Bayesian VAR models, prior distributions typically take up the following Gaussian form:[3]

$$\mathbf{a} \sim N\left(\mathbf{0}, \mathbf{I}\otimes\overline{\Sigma}\right), \tag{4}$$

$$\mathbf{a}_+|\mathbf{a} \sim N\left((\mathbf{I}\otimes\overline{\mathbf{P}})\mathbf{a}, \mathbf{I}\otimes\overline{\mathbf{H}}\right), \tag{5}$$

where $\overline{\Sigma}$ is an $n\times n$ symmetric, positive definite (SPD) matrix, $\overline{\mathbf{H}}$ is a $k\times k$ SPD matrix, and $\mathbf{P}$ is a $k\times n$ matrix. As shown in Sims and Zha (1997), the joint posterior p.d.f. of $(\mathbf{a},\mathbf{a}_+)$ is of form $p(\mathbf{a})\,p(\mathbf{a}|\mathbf{a}_+)$ in which

$$p(\mathbf{a}) \propto |\mathbf{A}|^T \exp\left(-\frac{1}{2} trace(\mathbf{A}'\mathbf{S}\mathbf{A})\right), \tag{6}$$

variables.

[3] See Sims and Zha (1997) for detailed discussion.

$$p(\mathbf{a}_+|\mathbf{a}) = \varphi\big((\mathbf{I}\otimes\mathbf{U})\mathbf{a};\mathbf{I}\otimes\mathbf{V}\big) \ , \tag{7}$$

where $\varphi(x;y)$ is a normal p.d.f. with mean x and covariance matrix y, $\mathbf{S}$ is a function of $(\mathbf{Y},\mathbf{X},\overline{\Sigma},\overline{\mathbf{H}})$, $\mathbf{U}$ is a function of $(\mathbf{Y},\mathbf{X},\overline{\mathbf{P}},\overline{\mathbf{H}})$, and $\mathbf{V}$ is a function of $(\mathbf{X},\overline{\mathbf{H}})$. Note that if there is no prior on $(\mathbf{a},\mathbf{a}_+)$ (i.e., a flat prior on $(\mathbf{a},\mathbf{a}_+)$), the joint posterior has the same form of distribution as (6) and (7) except $\mathbf{S}$, $\mathbf{U}$, and $\mathbf{V}$ are now functions of data $\mathbf{Y}$ and $\mathbf{X}$.[4]

In the identified VAR literature, some individual equations such as the "monetary policy equation" have clear economic interpretations. As in traditional simultaneous equations models, reversing signs of all coefficients in equations does not change any economic meanings. Furthermore, because reversing signs of coefficients in an equation is equivalent to the sign change of an identified shock in that equation, the interpretation of point-estimated impulse responses is never affected by these sign changes. To see this argument clearly, consider coefficients $\left(\mathbf{A}_{\cdot i}, \mathbf{A}_{+\cdot i}\right)$ in the i th equation of model (3), where subscript "$\cdot i$" represents the i th column of the matrix. To reverse signs of $\left(\mathbf{A}_{\cdot i}, \mathbf{A}_{+\cdot i}\right)$ is equivalent to post-multiplying system (3) by the diagonal matrix

$$\underset{n\times n}{\mathbf{R}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} ,$$

where "-1" is the element in the i th row and i th column. Impulse responses at time $s \geq 1$, denoted by $\underset{n\times n}{\Phi_s}$ in matrix form, are $\mathbf{A}^{-1}$ post-multiplied by an $n\times n$ matrix polynomial function of reduced-form parameter $\mathbf{B}$ $(=\mathbf{A}_{+}\mathbf{A}^{-1})$. Obviously, post-multiplying system (3) by $\mathbf{R}$ has no effect on the estimated value of $\mathbf{B}$. Thus, the new impulse responses, after the operation of a sign reserving, equal $\mathbf{R}\Phi_s$. These new responses to the i th shock, represented by the i th row of $\mathbf{R}\Phi_s$, simply flip the original impulse responses of variables to the i th shock around the zero axis. Suppose that the i th equation is identified as the "monetary policy reaction function" and that in response to this policy shock, the interest rate rises initially and the price level falls subsequently. These responses are interpreted as those to a *contractionary* policy shock. When signs of all estimated coefficients in this equation are reversed, all estimated impulse responses to this shock are accordingly flipped around the zero axis: the interest rate falls initially and the price level rises subsequently. The shock is now interpreted as an "*expansionary* policy shock" which is equivalent to reversing the sign of a "contractionary policy shock". Nothing is yet altered in regard to the essence of economic meanings except different labels or names are attached. Consequently, the sign of a particular equation can be fixed by arbitrarily restricting the value of one non-zero coefficient to be positive.

3. Normalization

The previous section articulates a well-known fact that estimation of identified multivariate model (3) does not depend on how each equation is normalized. In other words, the economic interpretation of the model, usually presented by the point estimates of impulse responses, is invariant to arbitrary rules of normalization. This argument, intuitive though it might seem to be, is no longer valid when a researcher desires to provide probability assessments for estimated impulse responses. Normalization can alter conclusions with respect to probability inferences. To show why this is true, let us follow the identified VAR literature to focus on linear

[4] See Sims and Zha (1997) and Zha (1997).

restrictions on contemporaneous coefficient matrix $\mathbf{A}$. Specifically, assume that there are q linear restrictions and each restriction applies to parameters in one equation. Denote the subspace of $\mathbf{R}^{n^2}$ with such q linear restrictions by

$$\mathcal{P} = \left\{ \mathbf{a} \in \mathbf{R}^{n^2} \middle| Q\mathbf{a} = \mathbf{0} \right\} ,$$

where Q is a $q \times n^2$ matrix and $\mathbf{0}$ is a $q \times 1$ vector of 0's. In the rest of this paper, the discussion about coefficient matrix $\mathbf{A}$ assumes that $\mathbf{a} \in \mathcal{P}$.[5]

To understand the shape of the likelihood, first consider marginal posterior distribution (6). Since $\mathbf{S}$ is positive definite, $p(\mathbf{a})$ tends to zero as $\|\mathbf{a}\|$ tends to infinity, where $\|\cdot\|$ denotes a usual Euclidean norm throughout this paper. Hence, by standard compactness arguments, $p(\mathbf{a})$ has a maximum. Let $\hat{\mathbf{a}}$ be a maximum point in $\mathcal{P}$. Since the maximum of conditional posterior p.d.f. (7) is the same for all $\mathbf{a}$, it must be true that $(\hat{\mathbf{a}}, (\mathbf{I} \otimes \mathbf{U})\hat{\mathbf{a}})$ is a maximum point for the joint posterior p.d.f. of $(\mathbf{a}, \mathbf{a}_+)$.

Now, consider reversing signs of coefficients in the i th equation. This implies reversing signs of all elements in the i th column of $\mathbf{A}$ as well as in the i th column of $\mathbf{A}_+$. From (6) and (7), it can be seen that sign changes in $\mathbf{a}_+$ are related to those in $\mathbf{a}$ through transformation $(\mathbf{I} \otimes \mathbf{U})\mathbf{a}$. Therefore, the value of the joint posterior p.d.f. of $(\mathbf{a}, \mathbf{a}_+)$ does not change when such sign reversing takes place. There are a total of 2^n possible sign changes. Consequently, the joint posterior probability distribution of $(\mathbf{A}, \mathbf{A}_+)$ is symmetric around the origin and across the 2^n subspaces.

Such symmetry makes normalization indispensable for probability inferences of the model's

[5] It is easy to see that the posterior p.d.f. of $\mathbf{a}$ in subspace $\mathcal{P}$ still has the same form as (6).

parameters. Without normalization, there would never exist a unique maximum likelihood (ML) estimate of $(\mathbf{A}, \mathbf{A}_+)$.[6] For instance, if $\hat{\mathbf{A}}$ is a maximum point of p.d.f. (6), a matrix obtained by reversing signs of one or more columns in $\hat{\mathbf{A}}$ is also a maximum point. Thus, there are always at least 2^n maximum points.

Since sign changes in $\mathbf{a}_+$ are related to those in $\mathbf{a}$ through transformation $(\mathbf{I} \otimes \mathbf{U})\mathbf{a}$, it is sufficient to discuss normalization only on coefficient vector $\mathbf{a}$ through this paper. To formalize the notion of normalization discussed so far, the following definition is in order.

Definition 1. For any given $\mathbf{a} \in \mathscr{P}$, $\mathcal{F}(\mathbf{a})$ is the set of all $\mathbf{b} \in \mathscr{P}$ such that $M(\mathbf{b})$ can be obtained by reversing signs of one or more columns of $M(\mathbf{a})$.

Note that there are a total of 2^n distinct elements in $\mathcal{F}(\mathbf{a})$. As all points in $\mathcal{F}(\mathbf{a})$ are identical up to sign changes, the essence of normalization is to summarize set $\mathcal{F}(\mathbf{a})$ by a single point. The idea of such normalization is now formalized by an intuitive operation as defined below.[7]

Definition 2. Normalization is a function g: $\mathscr{P} \rightarrow \mathscr{P}$ with the properties:

(1) $g(\mathbf{a}) \in \mathcal{F}(\mathbf{a})$;

(2) $\mathbf{b} \in \mathcal{F}(\mathbf{a})$ implies $g(\mathbf{b}) = g(\mathbf{a})$.

Clearly, normalization defined in Definition 2 can be intuitively thought of as collapsing set $\mathcal{F}(\mathbf{a})$ to single point $g(\mathbf{a})$. To see this, let $\mathcal{G}$ be the image of $\mathscr{P}$ under function *g*. Then the shape of posterior p.d.f. (6) on $\mathcal{G}$ provides all information about the shape of posterior (6)

[6] Following DeGroot (1970), this paper uses ML estimates to refer to generalized ML estimates which are maximum points of the joint posterior p.d.f. of $(\mathbf{A}, \mathbf{A}_+)$. Note that the likelihood is a posterior under a flat prior.

[7] For readers who are familiar with topology, note that Definition 2 is similar to the idea of using topological quotient spaces.

everywhere.

From the perspective of the likelihood principle, it is desirable to inform readers of the shape of posterior (6). This requires that normalization set the boundary of G farthest away from the peak of (6) to preserve the shape of the likelihood or the posterior p.d.f. Such normalization is called in this paper "ML distance normalization".

To set out a practical algorithm for carrying out ML distance normalization, a few notations and a definition are in order. Denote the columns of $M(\hat{\mathbf{a}})$ by $\hat{\mathbf{a}}_1,\ldots,\hat{\mathbf{a}}_n$ and the columns of $M(\mathbf{b})$ by $\mathbf{b}_1,\ldots,\mathbf{b}_n$, where $\hat{\mathbf{a}}$ is a ML estimate.

Definition 3. ML distance normalization is normalization g: $\mathscr{P} \to \mathscr{P}$ (as defined in Definition 2) with the property that for any point $\mathbf{b} \in \mathbf{R}^q$, $\|g(\mathbf{b}) - \hat{\mathbf{a}}\| \le \|\mathbf{b}' - \hat{\mathbf{a}}\|$ for $\mathbf{b}' \in \mathsf{F}(\mathbf{b})$.

With Definition 3, the following algorithm carries out ML distance normalization.

Algorithm 1. Moving from **b** to $g(\mathbf{b})$ involves three steps. For each i $(=1,\ldots,n)$,

(a) successively compute $\|\hat{\mathbf{a}}_j - \mathbf{b}_i\|$ and $\|\hat{\mathbf{a}}_j + \mathbf{b}_i\|$ for $j = i,\ i+1,\ \cdots,\ n,\ 1,\ \cdots,\ i-1$;

(b) stop at the first j such that $\|\hat{\mathbf{a}}_j + \mathbf{b}_i\| \ne \|\hat{\mathbf{a}}_j - \mathbf{b}_i\|$;

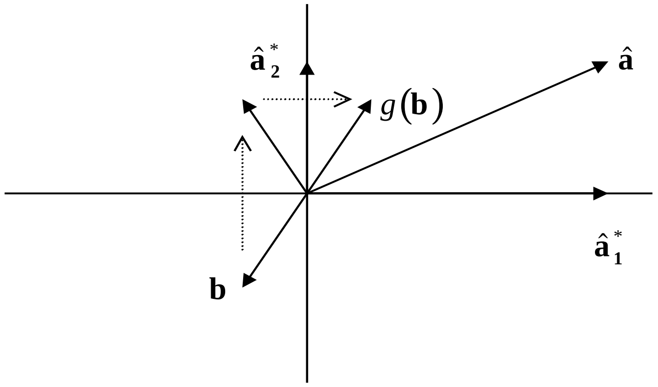
(c) replace $\mathbf{b}_i$ with $-\mathbf{b}_i$ if $\|\hat{\mathbf{a}}_j + \mathbf{b}_i\| < \|\hat{\mathbf{a}}_j - \mathbf{b}_i\|$ and leave $\mathbf{b}_i$ unchanged otherwise.

ML distance normalization $g(\mathbf{b})$ given by Algorithm 1 is well defined because there always exists stopping time j in step (b) for $\mathbf{b} \ne \mathbf{0}$. To see this point, suppose there does not exist a stopping time, i.e., $\|\hat{\mathbf{a}}_j + \mathbf{b}_i\| = \|\hat{\mathbf{a}}_j - \mathbf{b}_i\|$ for all j. Such a situation occurs only if $\mathbf{b}_i = \mathbf{0}$ because $M(\hat{\mathbf{a}})$ is non-singular. In this situation, it does not matter whether or not $\mathbf{b}_i$ is replaced by $-\mathbf{b}_i$.

Algorithm 1 ensures the mathematical property that ML distance normalization uniquely determines points on the boundary of G. These points may or may not belong to G after normalization. In practice, however, it is sufficient to consider only the interior of G because the set of points where $\|\hat{\mathbf{a}}_i + \mathbf{b}_i\| = \|\hat{\mathbf{a}}_i - \mathbf{b}_i\|$ for some i is a $\dim(\mathcal{P}) - 1$ dimensional subset of $\mathcal{P}$ and hence has measure zero. This property is important because, rather than finding the distance between 2^n points in $\mathbf{R}^q$ as required by Algorithm 1, one needs to compute the distance only between $2n$ points in $\mathbf{R}^q$. Specifically, a computationally efficient algorithm is simply to replace $\mathbf{b}_i$ with $-\mathbf{b}_i$ if $\|\hat{\mathbf{a}}_i + \mathbf{b}_i\| < \|\hat{\mathbf{a}}_i - \mathbf{b}_i\|$ for each $i = 1, \ldots, n$. Such an algorithm is valid because the distance from $\mathbf{b}$ to $\hat{\mathbf{a}}$ is $\sqrt{\sum_{i=1}^{n} \|\hat{\mathbf{a}}_i - \mathbf{b}_i\|^2}$.

At this point, it is instructive to present an example with $n = 2$ and Choleski identification. Let $\hat{\mathbf{a}} = \begin{bmatrix}\hat{\mathbf{a}}_1 & \hat{\mathbf{a}}_2\end{bmatrix}'$, $\hat{\mathbf{a}}_1^* = \begin{bmatrix}\hat{\mathbf{a}}_1 & \mathbf{0}\end{bmatrix}'$, and $\hat{\mathbf{a}}_2^* = \begin{bmatrix}\mathbf{0} & \hat{\mathbf{a}}_2\end{bmatrix}'$. Suppose restriction $\mathbf{a}_1(2) = 0$ is implied by Choleski identification. As a result, there are only three unrestricted parameters in $\mathbf{A}$. Clearly, $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$ are orthogonal to each other and lie on a hyperplane in $\mathbf{R}^3$ (here, $\mathcal{P} = \mathbf{R}^3$). First, consider a simple case in which $\mathbf{b}$ is in the linear span of $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$, which is pictured below.

Figure 1. ML Distance Normalization



The algorithm for ML distance normalization moves **b** along the dotted lines to element $g(\mathbf{b})$ that has the shortest distance from $\hat{\mathbf{a}}$. Now, consider a general case which $\mathbf{b}$ may not be lie completely in the span of $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$. In this case, decompose $\mathbf{b}$ into two components: the projected part and the perpendicular part. The projection of **b** onto the span of $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$ moves in the manner described in Figure 1; the perpendicular component of **b** to the span of $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$ remains all the time perpendicular to the span of $\hat{\mathbf{a}}_1^*$ and $\hat{\mathbf{a}}_2^*$. Thus, the interior of G is the set of all points in $\mathbf{R}^3$ that project onto the open first quadrant of the plane spanned by $\hat{\mathbf{a}}_1$ and $\hat{\mathbf{a}}_2$.

Traditional normalization used in the literature is to change signs of columns of $\mathbf{A}$ or equivalently signs of columns of $M(\mathbf{a})$ so that all diagonal elements are restricted to be positive. In light of Figure 1, it is easy to see that this normalization is equivalent to distance normalization described in Algorithm 1 but with ML estimate $\hat{\mathbf{a}}$ replaced by $vec(\mathbf{I})$. By moving $\mathbf{a}$ to $g(\mathbf{a})$ that is closest to $vec(\mathbf{I})$ rather than ML estimate $\hat{\mathbf{a}}$, however, traditional normalization generates G that is different from that generated by ML distance normalization. Thus, the implied shape of the likelihood or the posterior p.d.f. is different. The difference is likely to lead to quite different inferences about, say, impulse responses. To see this argument clearly, it may help to focus on two equations in an identified system.

Example 1. A Heuristic Case

Consider money supply (MS) and money demand (MD) equations of the form:

$$\begin{aligned} &MS: \quad a_1 R(t) + a_2 M(t) + \beta_s X_s(t) = \varepsilon_{MS}(t) \\ &MD: \quad a_3 R(t) + a_4 M(t) + \beta_d X_d(t) = \varepsilon_{MD}(t) \end{aligned}, \tag{8}$$

where R is the interest rate, M is the money stock, X_s contains all other variables in the MS equation, and X_d contains all other variables in the MD equation. For clear exposition, consider $\mathbf{a} = (a_1, a_2, a_3, a_4)'$ exclusively. Thus,

$$\mathbf{A} \equiv M(\mathbf{a}) = \begin{bmatrix} a_1 & a_3 \\ a_2 & a_4 \end{bmatrix}.$$

> If all other equations in the system are contemporaneously block recursive to the MS and MD equations in the sense of Zha (1997) (i.e., variables $R(t)$ and $M(t)$ do not enter other equations), the matrix of first-period impulse responses of M and R to MS and MD shocks is simply the inverse of $M(\mathbf{a})$.[8] That is to say,
>
> $$\Phi_1 = \frac{1}{(a_1 a_4 - a_2 a_3)^{-1}} \begin{bmatrix} a_4 & -a_3 \\ -a_2 & a_1 \end{bmatrix}$$
>
> Now suppose the maximum of the posterior p.d.f. occurs at, say, $\hat{a}_2 = \hat{a}_3 = 100$ and $\hat{a}_1 = \hat{a}_4 = 0.1$ with very high probability that $|a_2 a_3| >> |a_1 a_4|$. Furthermore, assume (i) the marginal posterior p.d.f.'s of both a_3 and $a_2 a_3$ tend to zero as (a_2, a_3) moves away from ML point $(\hat{a}_2, \hat{a}_3)$ toward zero and (ii) a_4 can be either positive or negative with equal probability.

If the normalization is to restrict all diagonal elements (a_1 and a_4) to be positive for every $\mathbf{a}$ drawn from the posterior, it could induce an *artificially* large standard error of a_3 so that one may infer that both ML coefficient $\hat{a}_3$ and the estimated first-period impulse response of money M to shock ε_{MS}, $\hat{\Phi}_1(1,2)$, are statistically "insignificant". But this is precisely not the inference one should make because the shape of the posterior, by assumption, is such that a_3 is very unlikely to be zero and ML coefficient $\hat{a}_3$ is sharply estimated, not "insignificant". On the other hand, ML distance normalization will by definition deliver the correct inference: both coefficient $\hat{a}_3$ and impulse response $\hat{\Phi}_1(1,2)$ are sharply estimated.

Although normalization in Definition 2 is a well-defined notion, there exists numerous rules

[8] For example, Gordon and Leeper (1994) make this block recursive assumption in their 7-variable identified VAR

or ways of normalization that are consistent with this definition. Depending on particular problems, they may or may not preserve the shape of the likelihood as intended by ML distance normalization. For instance, instead of normalizing on all diagonal elements of $\mathbf{A}$, one can normalize on some non-zero off-diagonal elements. In Example 1, this means that one alternative rule is to reverse the sign of the first column if $a_2 < 0$ and the sign of the second column if $a_3 < 0$. By Definition 2, this rule is equivalent to moving $\mathbf{a}$ to $g(\mathbf{a})$ that is closest to point $vec\left(\begin{vmatrix} 0 & 1 \\ 1 & 0 \end{vmatrix}\right)$. Such a rule is certainly sensible for the situation presented in Example 1 because it is likely to yield inferences about impulse responses that are not grossly at odds with inferences derived by ML distance normalization. In general, however, the rule that normalizes on off-diagonal elements may distort the shape of the likelihood.

Another example is the rule that normalizes on the diagonal of $\mathbf{A}^{-1}$ rather than $\mathbf{A}$ itself: if $\mathbf{A}^{-1}(i,i)$ < 0, reverse the sign of the i th column of $\mathbf{A}$.[9] The idea behind this rule is that researchers are sometimes concerned only with impulse responses.[10] If the a priori belief is that a contractionary money supply shock ought to raise the interest rate (R) initially, a "good" rule of normalization should restrict the first-period response of R to shock ε_{MS} to be always positive. In Example 1, this means to keep the value of $a_4 / \left| M(\mathbf{a}) \right|$ positive by appropriately reversing the sign of the first column of $M(\mathbf{a})$. Such a rule is valid normalization by Definition 2 because reversing signs of coefficients in the i th equation (i.e., the i th column of $\mathbf{A}$) is equivalent to flipping the impulse responses of variables to the i th shock (i.e., the i th row of $\mathbf{A}^{-1}$) around the zero axis. Obviously, unless $\mathbf{A}$ is restricted to be upper triangular (usually

model.

[9] The authors thank Chris Sims for this thoughtful suggestion.

called “Choleksi decomposition” in the literature), this rule is generally different from the rule that normalizes on the diagonal of $\mathbf{A}$ because it moves $\mathbf{a}$ to $g(\mathbf{a})$ in the manner that $vec(M^{-1}(g(\mathbf{a})))$ is closest to fixed point $vec(\mathbf{I})$. For the same reason that applies to the normalization on the diagonal of $\mathbf{A}$, however, this alternative rule of normalization may still mislead one to infer that the impulse response of M to a money supply shock in Example 1 is “insignificant.”

4. Monte Carlo Method and Results

The previous section defines the concept of normalization in the identified VAR framework and argues for ML distance normalization from the perspective of the likelihood principle. In this section, a numerical example is given to show that the two popular rules of normalization, one based on the diagonal of $\mathbf{A}$ and the other on the diagonal of $\mathbf{A}^{-1}$, can yield misleading inferences. Before proceeding with such an example, however, this section first develops an efficient Monte Carlo method for generating random samples of $\mathbf{a}$ from posterior p.d.f. (6).

The posterior p.d.f. of $\mathbf{a}$ in (6) is not of any standard distribution. In general, there is no way to draw $\mathbf{a}$ directly from this posterior except in some special cases.[11] A general method so far used in the literature is the importance-sampling technique originally recommended by Sims and Zha (1995). The basic idea is to approximate true posterior p.d.f. (6) with a Gaussian or t-distribution. Unfortunately, as the degree of simultaneity in model (3) increases, the form of posterior p.d.f. (6) tends to be very non-Gaussian in shape. As documented by Leeper, Sims, and Zha (1996, p. 37), “Gaussian approximations to this form are so bad that importance sampling is prohibitively inefficient.”

[10] Uhlig (1997) and Faust (1997) explore this idea in different contexts.
[11] See Zha (1997) for detailed discussion.

A wide variety of Monte Carlo (MC) methods, in particular Markov Chain simulation methods, have been discussed extensively in the recent literature (e.g., Geweke (1995) and Chib and Greenberg (1995)). One MC method is called a “random walk Metropolis algorithm” (Tierney (1994)). Given target distribution $p(\mathbf{a})$ in (6), a Metropolis algorithm generates a sequence of random samples $(\mathbf{a}^{(1)},\mathbf{a}^{(2)},\ldots)$ whose distributions converge to the target distribution. Each sequence can be viewed as a random walk whose distribution is (6). Unlike importance sampling in which the approximate distribution remains the same, approximate distributions in the Metropolis algorithm improve at each step in the simulation. The Metropolis algorithm developed in this paper is now described as follows.

Algorithm 2. Initialize arbitrary value $\mathbf{a}^{(0)}$ in $\mathbf{R}^q$. For $n=1,\ldots,N_1+N_2$,

(a) generate $\mathbf{z}$ from $h(\mathbf{z})$ and u from uniform distribution $U(0,1)$ where $h(\cdot)$ is a student-t p.d.f. with $(\mathbf{0},c\mathbf{S},\upsilon,q)$ in which c is a scale factor and υ the number of degrees of freedom;

(b) compute $\mathbf{a}=\mathbf{a}^{(n-1)}+\mathbf{z}$ and

$$J\left(\mathbf{a}^{(n-1)},\mathbf{a}\right)=\min\left\{\frac{p(\mathbf{a})}{p\left(\mathbf{a}^{(n-1)}\right)},1\right\} \ ;$$

(c) if $u\le J\left(\mathbf{a}^{(n-1)},\mathbf{a}\right)$, set $\mathbf{a}^{(n)}=\mathbf{a}$; else, set $\mathbf{a}^{(n)}=\mathbf{a}^{(n-1)}$;

(d) simulate sequence $\left\{\mathbf{a}^{(1)},\ldots,\mathbf{a}^{(N_1+N_2)}\right\}$ and but keep only the last N_2 values of the sequence.

According to Tierney (1994), Algorithm 2 generates a sequence of random samples whose distributions converge to target distribution (6). Intuitively, step (b) in Algorithm 2 can be thought of as a stochastic version of stepwise optimization: when the value of the p.d.f. increases, always step to climb; when the value decreases, only sometimes step down.

Algorithm 2 proves quite efficient for identified VAR models even when the shape of posterior (6) is very non-Gaussian and importance sampling becomes inefficient.

Figure 2 reports results of the impulse responses to a monetary policy shock from Sims (1986)'s overidentified six-variable VAR model. The six variables are the 3-month Treasure Bill rate (R), M1, GNP (y), GNP deflator (P), the unemployment rate (U), and gross domestic business investment (I). All variables are in logarithm except the interest rate and the unemployment rate which are in an expression that already is divided by 100. The model uses quarterly data with Sims's sample period 1948:1-1979:3. The time horizon for all impulse responses is 16 quarters. The identifying restrictions follow exactly what is called "second identification" in the original paper. The prior specification explores the basic idea expressed in the original paper but takes up the exact form as in Leeper, Sims, and Zha (1996)[12]. The middle line in Figure 2 is ML-estimated impulse responses, derived from ML estimates of $(\mathbf{a},\mathbf{a}_+)$. The two outer lines are .95 probability bands.[13] These bands are computed from 1.8 million MC samples by first drawing $\mathbf{a}$ with Algorithm 2 and then drawing $\mathbf{a}_+$ directly from conditional Gaussian distribution (7).[14]

The first column of graphics in Figure 2 displays probability bands generated by ML distance

[12] The prior used here is simply a reference prior which is not influential on the characteristics of impulse responses. Rather, it is designed to eliminate erratic sampling errors as the model becomes large and to reduce tendency of overfitting the data in dynamic multivariate models. See Sims and Zha (1997) for detailed discussion.

[13] Algorithm 2 is used to generate these bands. In step (a) of Algorithm 2, scale factor "c" is set at 0.25 and the degrees of freedom "υ" are set at 3. The proportion of random draw $\mathbf{a}^{(n)}$ at the n th simulation moving to new point $\mathbf{a}$, typically called "the value of jumping rule $J(\mathbf{a}^{(n-1)},\mathbf{a})$," is about 0.70.

[14] To monitor convergence, three parallel sequences are simulated with dispersed starting points. Each sequence has 750,000 random simulations of which first 150,000 draws are discarded to ensure convergence. As a result, there are a total of 1.8 million effective draws. Computing time is about 40 minutes for every 100,000 draws on Pentium II. Convergence criterion uses a measure called "potential reduction scale" constructed by Gelman et al (1995). Such a measure weights both the mean of the three *within*-sequence variances and the variance *between* the three means of sequences (see Gelman and Rubin (1992) and Gelman et al (1995) for details). For all parameters, potential reduction scale is almost 1 (below 1.002), which suggests a very high level of precision in simulations. Of course, for many practical problems, a much fewer number of draws (say, 100,000) are actually needed to achieve

normalization. It shows that monetary policy shocks generate both a liquidity effect (the interest rate rises initially and the money stock declines steadily) and a contractionary effect (output, price, and investment all fall and the unemployment rate rises for about one and a half years). These results imply that ML-estimated impulse responses are quite informative. Such statistical reliability of estimated impulse responses could be distorted by other normalization rules. The second column in Figure 2 displays results produced from the normalization rule that restricts the diagonal of $\mathbf{A}$ to be positive; the third column displays results generated from the normalization rule that restricts the diagonal of $\mathbf{A}^{-1}$ to be positive. Both columns imply that the estimated dynamic impact of monetary policy shocks is not useful or informative because almost all impulse responses are statistically “insignificant”. But the conclusion of “statistical insignificance” is really an artifact of inappropriate normalization rules. Although the model analyzed here is more complicated than Example 1, some insights presented in the discussion of Example 1 help explain why the normalization rules used for columns 2 and 3 of Figure 2 are at odds with the shape of the likelihood.

5. Conclusion

In a simultaneous equations framework, it is well known that reversing signs of coefficients in equations is simply an outcome of normalization that does not change the model’s economic interpretation. Traditional approaches to normalization are to restrict arbitrarily any non-zero ML estimate to be, say, positive. This paper argues that the shape of the likelihood or the posterior distribution could be distorted by inappropriate normalization rules. It discusses the concept of normalization in the context of dynamic multivariate models and introduces the method of ML distance normalization. Moreover, the paper develops a new Monte Carlo

reasonable accuracy in approximations to the target posterior distribution.

Bayesian algorithm for computing probability bands for impulse responses. An example in the existing literature is used to highlight ML distance normalization as a way of preserving the shape of the likelihood.

References

Bernanke, Ben S., Mark Gertler and Mark Watson, 1997. "Systematic Monetary Policy and the Effects of Oil Price Shocks," *Brookings Papers on Economic Activity 1, 91-142.*

Chib, Siddhartha and Edward Greenberg, 1995. "Understanding the Metropolis-Hastings Algorithm," *The American Statistician 49 (4)*, 327-335.

Faust, Jon, 1997. "The Robustness of Identified VAR Conclusions About Money," *manuscript,* Board of Governors of the Federal Reserve System.

Gelman, Andrew and Donald B. Rubin, 1992. "A Single Sequence from the Gibbs Sampler Gives a False of Security," in *Bayesian Statistics 4*, ed. J.M. Bernardo, J.O. Berger, A.P. Dawid, and A.F. Smith (New York: Oxford University Press), 625-631.

Gelman, Andrew, John B. Carlin, Hal S. Stern, and Donald B. Rubin, 1995. *Bayesian Data Analysis*, London: Cahpman & Hall.

Geweke, John, 1995. "Monte Carlo Simulation and Numerical Integration," in H. Amman, D. Kendrick and J. Rust (eds.), *Handbook of Computational Economics*, Amsterdam: North-Holland.

Gordon, David B. and Eric M. Leeper, 1994. "The Dynamic Impacts of Monetary Policy: An Exercise in Tentative Identification," *Journal of Political Economy 102,* 1228-1247.

Judge, George G., R. Carter Hill, William E. Griffiths, Helmut Lutkepohl, Tsoung-Chao Lee, 1995. *The Theory and Practice of Econometrics*, 2nd ed., New York: Wiley.

Leeper, Eric M., Christopher A. Sims, and Tao Zha, 1996. "What Does Monetary Policy Do?" *Brookings Papers On Economic Activity 2*, 1-63.

DeGroot, Morris H., 1970. *Optimal Statistical Decisions*, New York: McGraw-Hill Publishing Company.

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, 1995. "Error Bands for Impulse Responses," *Federal Reserve Bank of Atlanta Working Paper 95-6*.

______________________________, 1997. "Bayesian Methods for Dynamic Multivariate Models," forthcoming, *International Economic Review*.

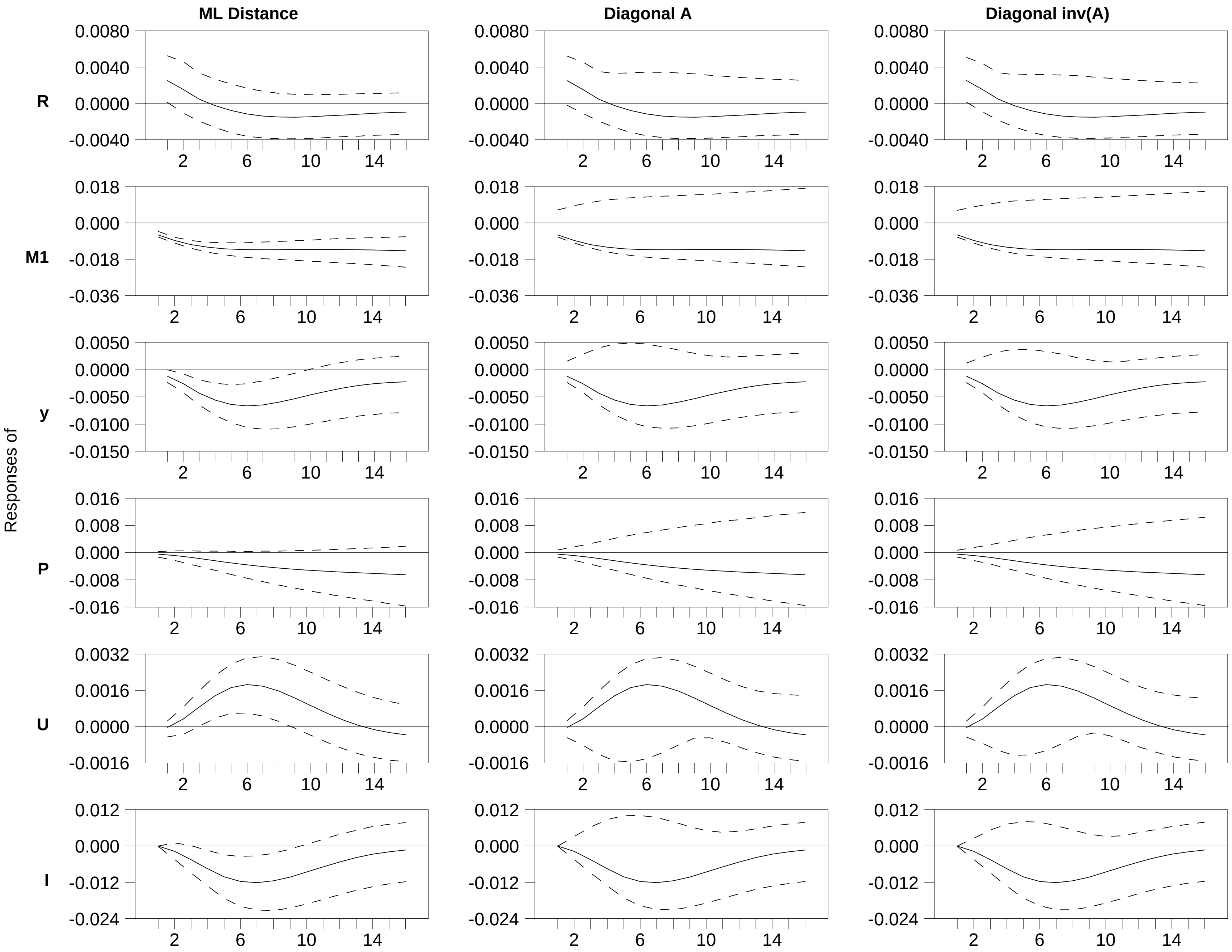
Tierney, L., 1994. "Markov Chains for Exploring Posterior Distributions," *Annals of Statistics 22*, 1701-1762.

Uhlig, Harald, 1997. "What Are the Effects of Monetary Policy? Results From an Agnostic

Identification Procedure," *manuscript*, Tilburg University.

Zha, Tao, 1997. "Block Recursion and Structural Vector Autoregressions," *manuscript*, Federal Reserve Bank of Atlanta.

Figure 2. Dynamic Responses to Monetary Policy Shock


ML Distance
Diagonal A
Diagonal inv(A)
Responses of
R
M1
y
P
U
I