Noninformative Priors and Frequentist Risks of Bayesian Estimators of Vector-Autoregressive Models

Shawn Ni, Department of Economics, University of Missouri, Columbia, MO 65211, USA

Dongchu Sun, Department of Statistics, University of Missouri, Columbia, MO 65211, USA

SUMMARY

In this study, we examine posterior properties and frequentist risks of Bayesian estimators based on several non-informative priors in Vector Autoregressive (VAR) models. We prove existence of the posterior distributions and posterior moments under a general class of priors. Using a variety of priors in this class we conduct numerical simulations of posteriors. We find that in most examples Bayesian estimators with a shrinkage prior on the VAR coefficients and the reference prior of Yang and Berger (1994) on the VAR covariance matrix dominate MLE, Bayesian estimators with the diffuse prior, and Bayesian estimators with the prior used in RATS. We also examine the informative Minnesota prior and find that its performance depends on the nature of the data sample and on the tightness of the Minnesota prior. A tightly set Minnesota prior is better when the data generating processes are similar to random walks, but the shrinkage prior or constant prior can be better otherwise.

KEY WORDS: VAR, Noninformative priors, constant prior, shrinkage prior, reference prior, Jeffreys prior, Minnesota prior.

JEL CLASSIFICATIONS: C11, C15, C32.

1 Introduction

Vector-Autoregression (VAR) models initiated by the seminal papers of Sims (1972, 1980) have become indispensable for macroeconomic research. A VAR of a p dimensional row-random vector $\mathbf{y}_t$, typically has the form

$$\mathbf{y}_t = \mathbf{c} + \sum_{i=1}^{L} \mathbf{y}_{t-i} \mathbf{B}_i + \boldsymbol{\epsilon}_t, \tag{1}$$

where $t = 1, \cdots, T$, $\mathbf{c}$ is a $1 \times p$ unknown vector, $\mathbf{B}_i$ $(i = 1, \cdots, L)$ is an unknown $p \times p$ matrix, $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_T$ are independently and identically distributed (iid) normal $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ errors, with a $p \times p$ unknown covariance matrix $\boldsymbol{\Sigma}$. We call L the lag of the VAR, and the $(Lp + 1) \times p$ unknown matrix $\boldsymbol{\Phi} = (\mathbf{c}', \mathbf{B}_1', \cdots, \mathbf{B}_L')'$ the regression coefficients. The VAR above imposes no restrictions on the coefficients $\boldsymbol{\Phi}$ and the covariance matrix $\boldsymbol{\Sigma}$. In applications, $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ can be estimated from time series macroeconomic data by Ordinary Least Square (OLS) or Maximum Likelihood Estimator (MLE). Accurate estimation of finite sample distributions of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is important for economic applications of the VAR model: In the recently developed structural VAR literature numerous authors (e.g. Sims (1986), Gordon and Leeper (1994), Sims and Zha (1998b), Pagan and Robertson (1998), Leeper and Zha (1999), and Lee and Ni (2002)) derive identification schemes based on the estimates of $\boldsymbol{\Sigma}$. Many quantities of interest such as point forecast or impulse responses are highly non-linear functions of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$, which makes it important to precisely estimate the VAR parameters. Unfortunately, the frequentist finite sample distributions of OLS (or ML) estimators of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ are unavailable. Asymptotic theory, on the other hand, may not be applicable for finite sample inferences of VARs for two reasons. First, a typical VAR model in macroeconomic research involves a large number of parameters, and the sample size of data is often not large enough to justify the use of asymptotic theory. Second, when nonlinear functions of the VAR coefficients are of interest, the asymptotic theory involves approximation of nonlinear functions, and the approximation becomes worse the more nonlinear the functions there are (see Kilian 1999).

An alternative to asymptotic theory is the Bayesian approach, which combines information from the sample and the prior to form a finite sample posterior distribution of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$. The present paper evaluates alternative Bayesian procedures in terms of frequentist risks for practitioners who are interested in finite sample distributions of VAR parameters.

The key element of Bayesian analysis is the choice of prior. The prior may be informative or noninformative. A commonly used informative prior for $\boldsymbol{\Phi}$ is the Minnesota prior (see Litterman 1986), which is a multivariate normal distribution. If researchers have justified beliefs about the hyper-parameters in the prior distributions, it is wise to use informative priors that reflect these

1

beliefs. But in practice, using informative prior has pitfalls. One problem is that prior information developed from experience may be irrelevant for a new data set. Another problem is that using informative priors makes comparing scientific reports more difficult. In many cases it is desirable to employ noninformative priors that offer much of the benefit of Bayesian analysis without the difficulties involved with using an informative prior.

Noninformative priors are designed to reflect the notion that a researcher has only vague knowledge about the distribution of the parameters of interest before he observes data. Alternative criteria may be used to reflect the vagueness of the researcher's knowledge. Thus there maybe more than one noninformative prior for a given problem. It is also true that the parametric form of a noninformative prior derived using a given criterion is problem-dependent. A recent review of various approaches for deriving noninformative priors can be found in Kass and Wasserman (1996).

For the covariance matrix Σ, a widely employed noninformative prior is the Jeffreys prior (Jeffreys, 1967). A modified version of the Jeffreys prior is put to use in RATS (Regression Analysis of Time Series, a software package popular among macroeconomists). This prior will be called the RATS prior hereafter. The Jeffreys prior is quite useful for single parameter problems but can be seriously deficient in multiparameter settings (see Berger and Bernardo, 1992). As alternatives, Berger and Bernardo's (1989, 1992) reference priors have been shown to be successful in various statistical models, especially for iid cases. One of the objectives of the present study is to examine the posterior of the VAR covariance matrix under these alternative priors.

In practice, researchers often combine separately derived priors for Φ and Σ as priors for (Φ, Σ). The most popular noninformative priors include the constant-Jeffreys prior, a constant prior for Φ and the Jeffreys prior for Σ, which was initially used for multivariate regression by Tiao and Zellner (1964) and Geisser (1965). This prior has been widely used in VAR models (e.g., see Kadiyala and Karlsson, 1997). In the RATS manual, the constant-RATS prior combination for (Φ, Σ) is used for generation of the standard error bands of impulse responses. Hundreds, if not thousands, of published empirical studies have used RATS-generated impulse responses and consequently this default prior. The constant prior, although is used quite often for VAR coefficients, is known to be inadmissible under quadratic loss for estimation of an unknown mean of vector with iid normal observations (see Berger and Strawderman, 1996). An alternative to the constant prior is a "shrinkage" prior for Φ, which has been used in estimating the unknown normal mean in iid cases (e.g, Baranchik, 1964), and in hierarchical linear mixed models (e.g., Berger and Strawderman, 1996). The shrinkage prior is a natural candidate for the VAR coefficients and will in this study be explored in the VAR setting.

The fact that all of the noninformative priors of (Φ, Σ) mentioned above are improper raises

a question on the propriety of the posterior distribution. [1] There exist situations in which the posterior is improper even though the full conditional distributions necessary for Markov Chain Monte Carlo (MCMC) simulations are all proper (e.g., Hobert and Casella 1996, Sun et al. 2001). Our first task in studying properties of VAR estimators under alternative priors is to show that the posteriors of $(\mathbf{\Phi}, \mathbf{\Sigma})$ under these priors are proper. We establish posterior propriety for a general class of priors that includes all prior combinations examined in the paper. In addition we also give proofs for existence of posterior moments. (The usefulness of the proofs is beyond the present paper.) Due to the fact that in most cases marginal posteriors are not available in closed-form, we use MCMC simulations to estimate posterior quantities numerically.

Besides comparing alternative noninformative priors, we also examine an informative Minnesota prior on $\mathbf{\Phi}$ used in combination with the reference prior on $\mathbf{\Sigma}$. The hyper-parameters for the Minnesota prior are defined as follows: The mean of $\mathbf{B}_1$ is set as the identity matrix, the mean of other elements in $\mathbf{\Phi}$ is set as zero, and the covariance matrix of the ith lag coefficient matrix $\mathbf{B}_i$ $(i = 1, 2, .., L)$ is set to be diagonal. In some macroeconomic applications, economic theory predicts random walk type behavior for certain time series (such as stock prices). If the true time series are random walks and the prior correctly reflects this information, then using an informative prior should lead to better performance. Intuitively, one would expect that when the Minnesota prior confirms to the VAR, it should do better for a tightly specified prior variance of $\mathbf{B}_i$. Otherwise the shrinkage prior or the constant prior may do better. We will investigate the performance of the priors when the data generating processes are of different types.

The rest of the paper is organized as follows. Section 2 lays out the notation and the MLE of the VAR model. Section 3 discusses the essential elements of Bayesian analysis for the VAR, including priors, posteriors, loss functions, and Bayesian estimators. Section 4 presents MCMC algorithms for Bayesian computation of posteriors. Section 5 reports numerical results of the Bayesian computation using noninformative priors. Finally, Section 6 presents some conclusions from this work.

2   Notations and the MLE of the VAR

We condiser the VAR model (1). Let

$$\mathbf{x}_t \quad = \quad (1, \mathbf{y}_{t-1}, \cdots, \mathbf{y}_{t-L}), \tag{2}$$

---

[1]A prior is improper if its integral over the entire parameter space is infinity.

$$\mathbf{Y} = \begin{pmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_T \end{pmatrix}, \ \mathbf{X} = \begin{pmatrix} \mathbf{x}_1 \\ \vdots \\ \mathbf{x}_T \end{pmatrix}, \boldsymbol{\epsilon} = \begin{pmatrix} \boldsymbol{\epsilon}_1 \\ \vdots \\ \boldsymbol{\epsilon}_T \end{pmatrix}, \boldsymbol{\Phi} = \begin{pmatrix} \mathbf{c} \\ \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_L \end{pmatrix}. \tag{3}$$

Here $\mathbf{Y}$ and $\boldsymbol{\epsilon}$ are $T \times p$ matrices, $\boldsymbol{\Phi}$ is a $(1 + Lp) \times p$ matrix of unknown parameters, $\mathbf{x}_t$ is a $1 \times (1 + Lp)$ row vector, and $\mathbf{X}$ is a $T \times (1 + Lp)$ matrix of observations. Then we rewrite (1) as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Phi} + \boldsymbol{\epsilon}. \tag{4}$$

The likelihood function of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is then

$$
\begin{aligned}
L(\boldsymbol{\Phi}, \boldsymbol{\Sigma}) &= \frac{1}{|\boldsymbol{\Sigma}|^{T/2}} \exp\left\{ -\frac{1}{2} \sum_{t=1}^{T} (\mathbf{y}_t - \mathbf{x}_t\boldsymbol{\Phi})\boldsymbol{\Sigma}^{-1}(\mathbf{y}_t - \mathbf{x}_t\boldsymbol{\Phi})' \right\} \\
&= \frac{1}{|\boldsymbol{\Sigma}|^{T/2}} \operatorname{etr}\left\{ -\frac{1}{2}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi})\boldsymbol{\Sigma}^{-1}(\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi})' \right\}.
\end{aligned}
\tag{5}
$$

Here and hereafter $\operatorname{etr}(\mathbf{A})$ is $\exp(\operatorname{trace}(\mathbf{A}))$ of a matrix $\mathbf{A}$. The finite sample distribution of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is the subject of interest. Note that the MLEs of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ are

$$\widehat{\boldsymbol{\Phi}}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y} \text{ and } \widehat{\boldsymbol{\Sigma}}_{MLE} = \mathbf{S}(\widehat{\boldsymbol{\Phi}}_{MLE})/T, \tag{6}$$

respectively, where

$$\mathbf{S}(\boldsymbol{\Phi}) = (\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi})'(\mathbf{Y} - \mathbf{X}\boldsymbol{\Phi}). \tag{7}$$

We assume that when $T \geq Lp+1$, $(\mathbf{X}'\mathbf{X})^{-1}$ exists with probability one, if $T \geq Lp+p+1$, $\mathbf{S}(\widehat{\boldsymbol{\Phi}}_{MLE})$ is positive definite, and the MLEs of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ exist with probability one. In this paper, we take as given that $T \geq Lp + p + 1$ so the MLEs of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ exist.

3 Bayesian Framework with Noninformative Priors

3.1 Priors for $\boldsymbol{\Phi}$

In practice, it is often convenient to consider vectorized VAR coefficients $\boldsymbol{\phi} = vec(\boldsymbol{\Phi})$, instead of $\boldsymbol{\Phi}$. A common expression of ignorance about $\boldsymbol{\phi}$ is a (flat) constant prior. For estimating the mean of a multivariate normal distribution, some authors (e.g., Baranchik 1964, Berger and Strawderman 1996) advocate the following "shrinkage" prior as an alternative to the constant prior for $\boldsymbol{\phi}$:

$$\pi_S(\boldsymbol{\phi}) \propto \|\boldsymbol{\phi}\|^{-(J-2)}, \quad \boldsymbol{\phi} \in \mathbb{R}^J, \tag{8}$$

where $J = p(Lp + 1)$, the dimension of $\boldsymbol{\phi}$. Berger and Strawderman show that the shrinkage prior (8) dominates the constant prior for estimating the iid normal means. The intuitive justification of

using the shrinkage prior on $\boldsymbol{\Phi}$ is related to the Stein (1956) effect, where the information about component variables can be used in such a way that "borrowed strength" improves the overall joint loss of the estimator. Berger and Strawderman make the following methodological recommandation on the choice of noninformative priors. "Avoid using constant priors for variances or covariance matrices, or for groups of mean parameters of dimension greater than 2." They add that "rigorous verifications of these recommendations would be difficult, but the results in this paper, together with our practical experience, suggest that they are very reasonable."

Our theoretical investigation on the posteriors is conducted in a framework that includes both the constant and shrinkage priors. We consider the class of priors of $\boldsymbol{\phi}$,

$$\pi_{(a)}(\boldsymbol{\phi}) \propto \frac{1}{\|\boldsymbol{\phi}\|^a}, \ a \geq 0. \tag{9}$$

When $a > 0$, $\pi_{(a)}(\boldsymbol{\phi})$ has the following two-stage hierarchical structure. Let $\pi_S(\boldsymbol{\phi}|\delta)$ be the normal density of $N_J(\mathbf{0}, \delta\mathbf{I}_J)$,

$$(\boldsymbol{\phi} \mid \delta) \ \sim \ N_J(\mathbf{0}, \delta\mathbf{I}_J) \ \text{ and assume } \ \pi_a(\delta) \ \propto \ \frac{1}{\delta^{\{a-(J-2)\}/2}}. \tag{10}$$

Then

$$\begin{aligned}
\int_0^\infty \pi_S(\boldsymbol{\phi} \mid \delta)\pi_{(a)}(\delta)d\delta &= \int_0^\infty \frac{1}{(2\pi\delta)^{J/2}} \exp\left\{-\frac{1}{2\delta}\boldsymbol{\phi}'\boldsymbol{\phi}\right\} \frac{1}{\delta^{\{a-(J-2)\}/2}}d\delta \\
&= \frac{1}{(2\pi)^{J/2}} \int_0^\infty \frac{1}{\delta^{a/2+1}} \exp\left\{-\frac{1}{2\delta}\boldsymbol{\phi}'\boldsymbol{\phi}\right\}d\delta \ = \ \frac{\Gamma(a/2)}{(2\pi)^{J/2}(\boldsymbol{\phi}'\boldsymbol{\phi})^{a/2}},
\end{aligned}$$

which is proportional to (9). As suggested in the introduction, informative priors are suitable vehicles for researchers to express their knowledge on the parameters of interest. A popular informative prior in macroeconomics is the so-called Minnesota prior on $\boldsymbol{\phi}$.

$$\pi_M(\boldsymbol{\phi}) \ \propto \ \frac{1}{|\mathbf{M}_0|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}_0)'\mathbf{M}_0^{-1}(\boldsymbol{\phi} - \boldsymbol{\phi}_0)\right\}. \tag{11}$$

In this paper we compare the Minnesota prior with the constant and shrinkage priors on $\boldsymbol{\phi}$. We will discuss the selection of hyper-parameters $\mathbf{M}_0$ and $\boldsymbol{\phi}_0$ later.

3.2 Priors for $\boldsymbol{\Sigma}$

The most popular noninformative prior for $\boldsymbol{\Sigma}$ is the Jeffreys prior (see Geisser 1965, Tiao and Zellner 1964). The Jeffreys prior is derived from the "invariance principle," meaning the prior is invariant to re-parameterization (see Zellner 1971). The Jeffreys prior is proportional to the square root of the determinant of the Fisher information matrix. Specifically, for the VAR covariance matrix, the Jeffreys prior is $\pi_J(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(p+1)/2}$. In RATS a modified version of the Jeffreys prior $\pi_A(\boldsymbol{\Sigma}) \propto |\boldsymbol{\Sigma}|^{-(L+1)p/2-1}$ is employed.

It has been noted, however, that the Jeffreys prior often gives unsatisfactory results for multi-parameter problems. For example, assuming the mean and variance are independent in the Neyman-Scott (1948) problem, the Bayesian estimator of the variance under the Jeffreys prior is inconsistent. An intuitive explanation for the poor performance of the Jeffreys prior in multi-parameter settings is that the parameter inter-dependence amplifies the effect of the prior on each parameter. Bernardo (1979) proposes an approach for deriving a reference prior by breaking a single multi-parameter problem into a consecutive series of problems with fewer numbers of parameters. The reference prior is designed to extract the maximum amount of expected information from the data in the sense of maximizing the difference (measured by Kullback-Leibler distance) between the posterior and the prior when the number of samples drawn goes to infinity. The reference priors preserve desirable features of the Jeffreys prior such as the invariance property, but they often avoid paradoxical results produced by Jeffreys prior in multi-parameter settings. Berger and Bernardo (1989, 1992) develop a procedure that leads to explicit forms of reference priors. They show that the Bayesian estimator of the variance in the Neyman-Scott problem is consistent under the reference prior. For other examples in which reference priors produce more desirable estimators than Jeffreys priors, see Berger and Bernardo (1992), Sun and Ye (1995), and Sun and Berger (1998), among others.

It is often the case that not all parameters deserve equal attention from a researcher. The key difference between the reference prior and the Jeffreys prior is that unlike the latter, the reference prior allows researchers to rank parameters by their perceived importance. For any given problem the reference prior depends on the ordering of the parameters. Bernardo (1979) shows that if the posterior is asymptotically normal, then the reference prior is the Jeffreys prior when there are no nuisance parameters. In estimating the variance-covariance matrix $\boldsymbol{\Sigma}$ based on an iid random sample from a normal population with known mean, Yang and Berger (1994) re-parameterize the matrix $\boldsymbol{\Sigma}$ as $\mathbf{O}'\mathbf{DO}$, where $\mathbf{D}$ is a diagonal matrix whose elements are the eigenvalues of $\boldsymbol{\Sigma}$ (in increasing or decreasing order), and $\mathbf{O}$ is an orthogonal matrix. The following reference prior is derived by giving priority to vectorized $\mathbf{D}$ over vectorized $\mathbf{O}$: $\pi_R(\boldsymbol{\Sigma}) \propto \{|\boldsymbol{\Sigma}|\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j)\}^{-1}$, where $\lambda_1 > \lambda_2 > \cdots > \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}$. Yang and Berger evaluate the reference-prior-based estimators of a covariance matrix in an iid setting.

Similar to our treatment of noninformative priors on $\boldsymbol{\Phi}$, we consider a general class of priors for $\boldsymbol{\Sigma}$,

$$\pi_{(b,c)}(\boldsymbol{\Sigma}) \propto \frac{1}{|\boldsymbol{\Sigma}|^{b/2}\{\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j)\}^c}, \tag{12}$$

where $b \in \mathbb{R}$ and $c = 0, 1$. Then $\pi_J(\boldsymbol{\Sigma})$, $\pi_A(\boldsymbol{\Sigma})$ and $\pi_R(\boldsymbol{\Sigma})$ are special cases of (12).

3.3 Joint Priors for $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$

The prior for $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ can be obtained by putting together priors for $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$. A popular noninformative prior for multivariate regression models is called the diffuse prior, which consists of a constant prior for $\boldsymbol{\phi}$ and the Jeffreys prior for $\boldsymbol{\Sigma}$. A similar prior is used in the RATS package. As will be shown later, the effect of the choice of prior for $\boldsymbol{\Phi}$ is not significantly affected by the prior on $\boldsymbol{\Sigma}$. For brevity, for evaluating the performance of the Minnesota prior, it suffices to report results of the Minnesota prior on $\boldsymbol{\phi}$ in combination with the reference prior on $\boldsymbol{\Sigma}$.

We now consider a general class of joint priors for $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$:

$$\pi_{(a,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \;\; = \;\; \pi_{(a)}(\boldsymbol{\phi})\pi_{(b,c)}(\boldsymbol{\Sigma}), \;\;\; c = 0, 1. \tag{13}$$

As special cases of (13), the prior combinations for $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ to be examined together with Minnesota-reference prior can be summarized as follows.

prior	notation	form	(a, b, c)
constant-Jeffreys	$\pi_{CJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{1}{|\boldsymbol{\Sigma}|^{(p+1)/2}}$	$(0, p+1, 0)$
constant-RATS	$\pi_{CA}(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$	$\frac{1}{|\boldsymbol{\Sigma}|^{(L+1)p/2+1}}$	$(0, (L+1)p+2, 0)$
constant-reference	$\pi_{CR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{1}{|\boldsymbol{\Sigma}|\prod_{1\le i<j\le p}(\lambda_i-\lambda_j)}$	$(0, 2, 1)$
shrinkage-Jeffreys	$\pi_{SJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{1}{\|\boldsymbol{\phi}\|^{J-2}|\boldsymbol{\Sigma}|^{(p+1)/2}}$	$(J-2, p+1, 0)$
shrinkage-RATS	$\pi_{SA}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{1}{\|\boldsymbol{\phi}\|^{J-2}|\boldsymbol{\Sigma}|^{(L+1)p/2+1}}$	$(J-2, (L+1)p+2, 0)$
shrinkage-reference	$\pi_{SR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{1}{\|\boldsymbol{\phi}\|^{J-2}|\boldsymbol{\Sigma}|\prod_{1\le i<j\le p}(\lambda_i-\lambda_j)}$	$(J-2, 2, 1)$
Minnesota-reference	$\pi_{MR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$	$\frac{\exp\left\{-\frac{1}{2}(\boldsymbol{\phi}-\boldsymbol{\phi}_0)'\mathbf{M}_0^{-1}(\boldsymbol{\phi}-\boldsymbol{\phi}_0)\right\}}{|\mathbf{M}_0|^{1/2}|\boldsymbol{\Sigma}|\prod_{1\le i<j\le p}(\lambda_i-\lambda_j)}$	

The list of noninformative priors examined in the present paper is by no means exhaustive. Other noninformative priors, such as Zellner's (1997) Maximal Data Information Prior (MDIP), can be derived using approaches not discussed in this paper. A modified version of Zellner's prior in a VAR setting is studied by Deschamps (2000). Sims and Zha (1998a) propose an MCMC procedure drawing $\boldsymbol{\Sigma}$ from an Inverse Wishart distribution and applying priors similar to the Minnesota prior on $\boldsymbol{\phi}$. The Sims-Zha approach is particularly convenient for estimation of identified VARs.

3.4 Propriety of the Posteriors

In this paper, we will compare various properties of estimators of the VAR parameters $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ under various noninformative priors. Since all the noninformative priors for $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ listed above are improper, it is important to know if the posteriors of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ exist under these priors. Sun and Ni (2001) prove that the posteriors of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ are proper under both the constant-Jeffreys and constant-reference priors $\pi_{CJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ and $\pi_{CR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. We now develop more general results on the posteriors under the prior $\pi_{(0,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$.

Theorem 1   *Consider the prior $\pi_{(0,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$.*
(a) If $T > (L+2)p - b + 1$, the posterior of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ under the prior $\pi_{(0,b,0)}$ is proper.
(b) If $T > Lp - b + 3 > 0$, the posterior of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ under the prior $\pi_{(0,b,1)}$ is proper.

The proof of the theorem is given in Appendix A. The next theorem shows that if the MLE exists, then the requirements on the sample size for existence of proper posteriors are satisfied for prior combinations involving the constant prior.

Theorem 2   *If the MLE of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ exists, then the posterior of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ is proper under $\pi_{CJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$, $\pi_{CA}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$, and $\pi_{CR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$.*

Proof.   In part (a) of Theorem 1, let $b = p + 1$ for prior π_{CJ} and $b = (L+1)p + 2$ for prior π_{CA}. The sample size requirement under $\pi_{CJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ is $T > (L+1)p$, and the requirement under $\pi_{CA}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ is $T > p - 1$. In part (b) of Theorem 1 with $b = 2$, the requirement under $\pi_{CR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ is $T > Lp + 1$. Existence of the MLE requires $T > (L+1)p + 1$, which guarantees the existence of the posterior under all three prior combinations. ☐

Theorem 2 implies that the posterior under the Minnesota-reference prior is proper due to the facts that the constant-reference prior is proper and the Minnesota prior is bounded by a constant.

To show the existence of the posterior under the prior $\pi_{(a,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ when $a > 0$, we introduce the following conditions.

(A) $J - a > 0$.

(B0) $T > \max(2p - b, J - a - b + 2)$.

(B1) $T > J - a - b + 2$.

Theorem 3   *Consider the prior $\pi_{(a,b,c)}$ when $a > 0$.*
(a) If Conditions (A) and (B0) hold, the posterior of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ under the prior $\pi_{(a,b,0)}$ is proper.
(b) If Conditions (A) and (B1) hold, the posterior of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ under the prior $\pi_{(a,b,1)}$ is proper.

The proof of the theorem is given in Appendix B.

Theorem 4 *If the MLE of $(\mathbf{\Phi}, \mathbf{\Sigma})$ exists, then posterior of $(\boldsymbol{\phi}, \mathbf{\Sigma})$ is proper under $\pi_{SJ}(\boldsymbol{\phi}, \mathbf{\Sigma})$, $\pi_{SA}(\boldsymbol{\phi}, \mathbf{\Sigma})$, and $\pi_{SR}(\boldsymbol{\phi}, \mathbf{\Sigma})$.*

Proof. Under prior π_{SJ}, applying part (a) of Theorem 3 with $a = J - 2$ and $b = p + 1$ leads to the requirement $T > max(p - 1, 3 - p)$. Under prior π_{SA} applying part (a) of Theorem 3 with $a = J - 2$ and $b = (L + 1)p + 2$ leads to the requirement $T > 2 - (L + 1)p$. Under prior π_{SR}, letting $a = J - 2$ and $b = 2$ in part (b) of Theorem 3 leads to the requirement $T > 2$. These requirements are satisfied if the MLE exists. □

3.5   Existence of Posterior Moments

Computing Bayesian estimators of VAR models involves posterior moments of $(\mathbf{\Phi}, \mathbf{\Sigma})$. Existence of the posterior is necessary but not sufficient for existence of posterior moments. In the following, we derive sufficient conditions for existence of posterior moments of certain orders. We first consider the case $a = 0$.

Theorem 5 *(a) If $T > (L + 2)p + 2h - b + 3$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ under the prior $\pi_{(0,b,0)}$ is finite, where h is a nonnegative integer.*
(b) If $T > Lp + 2h - b + 5$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ under the prior $\pi_{(0,b,1)}$ is finite.

The proof of the theorem is given in Appendix C. The results imply the existence of the first two posterior moments of the components of $\boldsymbol{\phi}$, and the hth posterior moments of the components of $\mathbf{\Sigma}$. The following theorem for the priors considered in this paper is a straightforward application of Theorem 5.

Theorem 6 *(a) Under $\pi_{CJ}(\boldsymbol{\phi}, \mathbf{\Sigma})$, if $T > (L+1)p+2+2h$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ is finite.*
(b) Under $\pi_{CA}(\boldsymbol{\phi}, \mathbf{\Sigma})$, if $T > p + 2h + 1$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ is finite.
(c) Under $\pi_{CR}(\boldsymbol{\phi}, \mathbf{\Sigma})$, if $T > Lp + 1$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ is finite.

Following part (c) of the theorem above, under $\pi_{MR}(\boldsymbol{\phi}, \mathbf{\Sigma})$, the posterior mean of $\|\boldsymbol{\phi}\|^2\{tr(\mathbf{\Sigma}^2)\}^{h/2}$ exists if $T > Lp + 1$.

Let k and h be nonnegative integers. Consider the conditions for the case $a > 0$:

(AM) $J - a > 0$ and $a - k > 0$;

9

(B0M) $T > \max(2p - b + 2h, J - a + k - b + 2)$;

(B1M) $T > J - a + k - b + 2$.

Theorem 7 (a) If Conditions (AM) and (B0M) hold, the posterior mean of $\|\boldsymbol{\phi}\|^k \{tr(\boldsymbol{\Sigma}^2)\}^{h/2}$ under the prior $\pi_{(a,b,0)}$ is finite.
(b) If Conditions (AM) and (B1M) hold, the posterior mean of $\|\boldsymbol{\phi}\|^k \{tr(\boldsymbol{\Sigma}^2)\}^{h/2}$ under the prior $\pi_{(a,b,1)}$ is finite.

The proof of the theorem is given in Appendix D. The results imply the existence of the kth posterior moments of the components of $\boldsymbol{\phi}$ and the hth posterior moments of the components of $\boldsymbol{\Sigma}$. Applying Theorem 7 to prior combinations that involve the shrinkage prior, we have the following result.

Theorem 8 (a) Under $\pi_{SJ}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$, if $T > \max(p - 1 + 2h, 3 - p + k)$, the posterior mean of $\|\boldsymbol{\phi}\|^k \{tr(\boldsymbol{\Sigma}^2)\}^{h/2}$ is finite.
(b) Under $\pi_{SA}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$, if $T > \max(p - Lp - 2 + 2h, k - (L+1)p + 2)$, the posterior mean of $\|\boldsymbol{\phi}\|^k \{tr(\boldsymbol{\Sigma}^2)\}^{h/2}$ is finite.
(c) Under $\pi_{SR}(\boldsymbol{\phi}, \boldsymbol{\Sigma})$, if $T > 2 + k$, the posterior mean of $\|\boldsymbol{\phi}\|^k \{tr(\boldsymbol{\Sigma}^2)\}^{h/2}$ is finite.

From Theorems 6 and 8 we conclude that the requirements on the sample size for existence of posterior moments are easily satisfied in practical cases.

3.6 Conditional Posterior Distributions

The posteriors of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ are not available in closed-form for most prior combinations. Recent years have witnessed vast progress in numerical posterior simulations. For some recent examples of Bayesian computations in econometrics, see Geweke (1996, 1999), Chib (1998), Chib and Hamilton (2000), and the references therein. In this study, we use Gibbs sampling MCMC methods to sample from the posteriors (cf. Gelfand and Smith, 1990). The first step of the MCMC computation is to find the full conditional distributions of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. We will make use of the following results.

Fact 1 Consider the constant-Jeffreys prior for $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. The conditional posterior of $\boldsymbol{\phi}$ given $\boldsymbol{\Sigma}$ is $MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \boldsymbol{\Sigma} \otimes (\boldsymbol{X'X})^{-1})$ and the marginal posterior of $\boldsymbol{\Sigma}$ is Inverse Wishart $(\boldsymbol{S}(\widehat{\boldsymbol{\Phi}}_{MLE}), T - Lp - 1)$. Here $\widehat{\boldsymbol{\phi}}_{MLE}$ is defined as vectorized $\widehat{\boldsymbol{\Phi}}_{MLE}$ and $\otimes$ is Kronecker product.

Proof. This follows from the proof of Theorem 1. (We followed the notation of the Inverse Wishart distribution of Anderson 1984, p268). □

Fact 2 *Consider the constant-RATS prior for* $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. *The conditional posterior of* $\boldsymbol{\phi}$ *given* $\boldsymbol{\Sigma}$ *is* $MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1})$ *and the marginal posterior of* $\boldsymbol{\Sigma}$ *is Inverse Wishart* $(\boldsymbol{S}(\widehat{\boldsymbol{\Phi}}_{MLE}), T)$.

Fact 3 *Consider the constant-reference prior.*

(a) The conditional distribution of $\boldsymbol{\phi}$ *given* $(\boldsymbol{\Sigma}, \boldsymbol{Y})$ *is*

$$\pi(\boldsymbol{\phi} \mid \boldsymbol{\Sigma}, \boldsymbol{Y}) \sim MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \ \boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1}). \tag{14}$$

(b) The conditional density of $\boldsymbol{\Sigma}$ *given* $(\boldsymbol{\phi}, \boldsymbol{Y})$ *is*

$$\pi(\boldsymbol{\Sigma} \mid \boldsymbol{\phi}, \boldsymbol{Y}) \propto \frac{etr\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\boldsymbol{S}(\boldsymbol{\phi})\}}{|\boldsymbol{\Sigma}|^{\frac{T}{2}+1} \displaystyle\prod_{1 \le i < j \le p} (\lambda_i - \lambda_j)}, \tag{15}$$

where $\boldsymbol{S}(\boldsymbol{\Phi})$ *is defined by (7).*

Proof. This follows from standard computation. □

The hierarchical structure (10) suggests a nice computational formula. For example, the shrinkage prior $\pi_S(\boldsymbol{\phi})$ is a special case of (10) with $a = J - 2$. In this case, we have

$$(\boldsymbol{\phi} \mid \delta) \ \sim \ N_J(\boldsymbol{0}, \delta \mathbf{I}_J) \ \text{ and } \ \pi(\delta) \ \propto \ 1.$$

Instead of simulating from the conditional distribution of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ within each Gibbs cycle, we use δ as a latent variable and simulate from $\boldsymbol{\Phi}$, $\boldsymbol{\Sigma}$, and δ based on the following fact.

Fact 4 *Consider the shrinkage-reference prior.*

(a) The conditional density of $\boldsymbol{\Sigma}$ *given* $(\boldsymbol{\phi}, \delta, \boldsymbol{Y})$ *is in (15).*

(b) The conditional distribution of $\boldsymbol{\phi} = vec(\boldsymbol{\Phi})$ *given* $(\delta, \boldsymbol{\Sigma}, \boldsymbol{Y})$ *is* $MVN_J(\boldsymbol{\mu}_S, \boldsymbol{V}_S)$, *where*

$$\boldsymbol{\mu}_S \ = \ \delta\left(\boldsymbol{\Sigma} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1} \ + \ \delta\boldsymbol{I}_J\right)^{-1}\widehat{\boldsymbol{\phi}}_{MLE}; \tag{16}$$

$$\boldsymbol{V}_S \ = \ \left(\boldsymbol{\Sigma}^{-1} \otimes \boldsymbol{X}'\boldsymbol{X} \ + \ \frac{1}{\delta}\boldsymbol{I}_J\right)^{-1}. \tag{17}$$

(c) The conditional distribution of δ *given* $(\boldsymbol{\Phi}, \boldsymbol{\Sigma}, \boldsymbol{Y})$ *is Inverse Gamma* $(\frac{J}{2} - 1, \frac{1}{2}\boldsymbol{\phi}'\boldsymbol{\phi})$.

Proof. The proof of (b) is similar to Example 9 of Berger (1984). The others are simple. □

Since the Minnesota prior of $\boldsymbol{\Phi}$ is independent of $\boldsymbol{\Sigma}$, the conditional posterior density under the Minnesota-reference prior for $\boldsymbol{\Sigma}$ given $(\boldsymbol{\phi}, \mathbf{Y})$ is given by (15). The conditional posterior density of $\boldsymbol{\phi}$ given $(\boldsymbol{\Sigma}, \mathbf{Y})$ is

$$\pi(\boldsymbol{\phi} \mid \boldsymbol{\Sigma}, \mathbf{Y}) \propto \exp\left\{-\frac{1}{2}(\boldsymbol{\phi} - \boldsymbol{\phi}_0)' M_0^{-1}(\boldsymbol{\phi} - \boldsymbol{\phi}_0) - \frac{1}{2}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}_{MLE})'\left[\boldsymbol{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})\right](\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}_{MLE})\right\}. \tag{18}$$

Thus we have the following result.

Fact 5 *Consider the Minnesota-reference prior. The conditional density of $\boldsymbol{\phi}$ given $(\boldsymbol{\Sigma}, \boldsymbol{Y})$ is $MVN_J(\boldsymbol{\mu}_M, \boldsymbol{V}_M)$, where*

$$\boldsymbol{\mu}_M = \widehat{\boldsymbol{\phi}}_{MLE} + (M_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1})^{-1} M_0^{-1}(\boldsymbol{\phi}_0 - \widehat{\boldsymbol{\phi}}_{MLE}); \tag{19}$$

$$\boldsymbol{V_M} = (M_0^{-1} + \boldsymbol{\Sigma}^{-1} \otimes (\boldsymbol{X}'\boldsymbol{X}))^{-1}. \tag{20}$$

The hyper-parameter $\boldsymbol{\phi}_0$ (i.e. $vec(\boldsymbol{\Phi}_0)$) is defined by letting the mean of $\mathbf{B_1}$ be the identity matrix and the mean of the other elements be zero. The elements of M_0 are given as b_1/k for parameters of VAR variables of their own kth lag; $(b_1 b_2/k)\,(\widehat{\sigma}_i/\widehat{\sigma}_j)^{1/2}$ for parameters of the kth lag of the jth variable in the ith equation, $j \neq i$, ($\widehat{\sigma}_i$ is the variance of the residuals of ith VAR equation estimated via OLS); and b_3 for intercepts. There is no unique way of choosing the hyper-parameters. Our specification of the M_0 matrix closely follows that of Kadiyala and Karlsson (1997), which is slightly different from the form in the RATS manual and Hamilton's book (1994 p361-362). In our numerical examples, we experiment with alternative settings of the Minnesota prior with different variance parameters b_1 and b_3. Following convention we choose the hyper-parameter b_2 to be 0.5. We also set $b_3 = 1.0$. We control the "tightness" of the Minnesota prior by adjusting the values of parameter b_1. A tight version of the Minnesota prior is defined by $b_1 = 0.2^2$, and a loose version sets $b_1 = 0.9^2$. Here the words "tight" or "loose" are used in relative terms. One can certainly argue that $b_1 = 0.9^2$ represents a tight prior compared to the case $b_1 = 10^2$.

3.7 Loss Functions and Bayesian Estimators

A Bayesian estimator of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ depends on the data generating model, the prior, and the loss function. We consider a pseudo entropy loss function for $\boldsymbol{\Sigma}$ and a quadratic loss function for $\boldsymbol{\Phi}$,

$$L_1(\widehat{\boldsymbol{\Sigma}}; \boldsymbol{\Sigma}) = \text{tr}(\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}) - \log|\widehat{\boldsymbol{\Sigma}}^{-1}\boldsymbol{\Sigma}| - p; \tag{21}$$

$$L_2(\widehat{\boldsymbol{\Phi}}, \boldsymbol{\Phi}) = \text{tr}\{(\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi})'\mathbf{W}^{-1}(\widehat{\boldsymbol{\Phi}} - \boldsymbol{\Phi})\}, \tag{22}$$

where $\mathbf{W}^{-1}$ is a constant weighting matrix, and p is the number of variables in the VAR. If the weighting matrix $\mathbf{W}$ is the identity matrix, the loss L_2 is simply the sum of squared errors of all elements of $\widehat{\boldsymbol{\Phi}}$,

$$\sum_{i=1}^{1+Lp} \sum_{j=1}^{p} (\widehat{\Phi}_{i,j} - \Phi_{i,j})^2. \tag{23}$$

The loss L_2 can be decomposed as $L_{21} + L_{22}$, where the loss associated with the intercept terms is

$$L_{21} = \sum_{j=1}^{p} (\widehat{\Phi}_{1,j} - \Phi_{1,j})^2, \tag{24}$$

and the loss associated with terms other than the intercepts is

$$L_{22} = \sum_{i=2}^{1+Lp} \sum_{j=1}^{p} (\widehat{\Phi}_{i,j} - \Phi_{i,j})^2. \tag{25}$$

The loss function for $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ contains a part measuring the loss associated with the covariance matrix (L_1) and a part measuring the loss pertaining to the VAR coefficients (L_2). It is well known that the Bayes estimator under the square error loss is the posterior mean. One can also verify that the posterior mean is the Bayesian estimator under loss function L_1. Thus we have the following result.

Lemma 1   *Under the loss $L_1 + L_2$, the generalized Bayesian estimator of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ is*

$$\widehat{\boldsymbol{\Phi}} = \mathbb{E}(\boldsymbol{\Phi} \mid \boldsymbol{Y}), \tag{26}$$

$$\widehat{\boldsymbol{\Sigma}} = \mathbb{E}(\boldsymbol{\Sigma} \mid \boldsymbol{Y}). \tag{27}$$

4   Algorithms for Simulating from Posterior of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$

The algorithms for MCMC computations of posterior distributions of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ depend on the priors. For brevity we only outline the algorithms with constant prior on $\boldsymbol{\phi}$ and the Jeffreys and reference priors on $\boldsymbol{\Sigma}$.

Following Fact 1, we use an MC algorithm to sample from the joint posterior distribution $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$. Suppose at cycle k we have $(\boldsymbol{\Phi_{k-1}}, \boldsymbol{\Sigma}_{k-1})$ sampled from cycle $k - 1$. The following algorithm is used for computing the posterior under the constant-Jeffreys prior.

Algorithm CJ:

Step 1: Simulate $\boldsymbol{\Omega} \sim IW(\mathbf{S}(\widehat{\boldsymbol{\Phi}}_{MLE}), T - Lp - 1)$ and let $\boldsymbol{\Sigma}_k = \boldsymbol{\Omega}$.

Step 2: Simulate $\boldsymbol{\phi}_k$ from $MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \boldsymbol{\Sigma}_k \otimes (\mathbf{X}'\mathbf{X})^{-1})$. Stop if $k + 1$ is larger than a pre-set number M, otherwise replace k by $k + 1$ and go to Step 1.

The algorithm using the constant-RATS prior is similar to the one above, with the exception that in Step 1 the distribution of the Inverse Wishart has different degrees of freedom: $\boldsymbol{\Omega} \sim IW(\mathbf{S}(\widehat{\boldsymbol{\Phi}}), T)$.

It is much more difficult to simulate from the conditional distribution of $\boldsymbol{\Sigma}$ under the reference prior. We adopt a hit-and-run algorithm used in Yang and Berger (1994). In implementing the algorithm, we consider a one-to-one transformation of $\boldsymbol{\Sigma}$, namely $\boldsymbol{\Sigma}^* = \log(\boldsymbol{\Sigma})$ or $\boldsymbol{\Sigma} = \exp(\boldsymbol{\Sigma}^*)$ in the sense that

$$\boldsymbol{\Sigma} = \sum_{j=0}^{\infty} \frac{\boldsymbol{\Sigma}^{*j}}{j!}.$$

13

It can be shown that the conditional posterior density of $\mathbf{\Sigma}^*$ given $(\boldsymbol{\phi}, \mathbf{Y})$ is then

$$\pi(\mathbf{\Sigma}^* \mid \boldsymbol{\phi}, \mathbf{Y}) = \pi(\mathbf{\Sigma}^* \mid \mathbf{S}(\boldsymbol{\phi})) \quad \propto \quad \frac{\text{etr}\{-\frac{T}{2}|\mathbf{\Sigma}^*| - \frac{1}{2}(\exp\mathbf{\Sigma}^*)^{-1}\mathbf{S}(\boldsymbol{\phi})\}}{\prod_{i<j}(\lambda_i^* - \lambda_j^*)}, \tag{28}$$

where $\mathbf{\Sigma}^* = \mathbf{O}'\mathbf{\Lambda}^*\mathbf{O}$, $\mathbf{O}$ is an orthogonal matrix, and $\mathbf{\Lambda}^* = diag(\lambda_1^*, \cdots, \lambda_p^*)$ with $\lambda_1^* > \ldots > \lambda_p^*$. Note that $\exp(\mathbf{\Sigma}^*) = \mathbf{O}'\exp(\mathbf{\Lambda}^*)\mathbf{O}$.

To simulate $\mathbf{\Sigma}^*$ from (28), we use the following algorithm. Assume we have a Gibbs sample $(\mathbf{\Phi}_{k-1}, \mathbf{\Sigma}_{k-1})$.

For Cycle k:

Step 1: Simulate $\boldsymbol{\phi}_k \sim MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \; \mathbf{\Sigma}_{k-1} \otimes (\mathbf{X}'\mathbf{X})^{-1})$ and get $\mathbf{\Phi}_k$.

Step 2: Calculate $\mathbf{S}_k = S(\boldsymbol{\phi}_k) = (\mathbf{Y} - \mathbf{X}\mathbf{\Phi}_k)'(\mathbf{Y} - \mathbf{X}\mathbf{\Phi}_k)$.

Step 3: Decompose $\mathbf{\Sigma}_{k-1} = \mathbf{O}'\mathbf{\Lambda}\mathbf{O}$, where $\mathbf{\Lambda} = diag(\lambda_1, \cdots, \lambda_p)$, $\lambda_1 > \lambda_2 \cdots > \lambda_p$, and $\mathbf{O}'\mathbf{O} = \mathbf{I}$. Let $\lambda_i^* = \log(\lambda_i)$, $\mathbf{\Lambda}^* = diag(\lambda_1^*, \cdots, \lambda_p^*)$, and $\mathbf{\Sigma}_{k-1}^* = \mathbf{O}\mathbf{\Lambda}^*\mathbf{O}'$.

Step 4: Select a random symmetric $p \times p$ matrix $\mathbf{V}$, with elements $v_{ij} = z_{ij}/\sqrt{\sum_{l \leq m} z_{lm}^2}$, where $z_{ij} \sim N(0,1)$, $1 \leq i \leq j \leq p$. The other elements of $\mathbf{V}$ are defined by symmetry.

Step 5: Generate $t \sim N(0,1)$ and set $\mathbf{W} = \mathbf{\Sigma}_{k-1}^* + t\mathbf{V}$. Decompose $\mathbf{W} = \mathbf{Q}'\mathbf{C}^*\mathbf{Q}$, where $\mathbf{C}^* = diag(c_1^*, \cdots, c_p^*)$, $c_1^* > c_2^* \cdots > c_p^*$, and $\mathbf{Q}'\mathbf{Q} = \mathbf{I}$. Compute

$$\begin{aligned}
\alpha_k &= \log(\pi(\exp(\mathbf{W}) \mid \mathbf{S}_k)) - \log(\pi(\exp(\mathbf{\Sigma}_{k-1}^*)) \mid \mathbf{S}_k) \\
&= \frac{T}{2}\sum_{i=1}^{p}(\lambda_i^* - c_i^*) + \frac{1}{2}\text{tr}\{((\exp\mathbf{\Sigma}_{k-1}^*)^{-1} - (\exp\mathbf{W})^{-1})\mathbf{S}_k\} \\
&\quad + \sum_{i<j}\log(\lambda_i^* - \lambda_j^*) - \sum_{i<j}\log(c_i^* - c_j^*).
\end{aligned}$$

Step 6: Generate $u \sim Unif(0,1)$.

If $u \leq min(1, \exp(\alpha_k))$, let $\mathbf{\Sigma}_k^* = \mathbf{W}$ and $\mathbf{\Sigma}_k = \mathbf{Q}\mathbf{C}\mathbf{Q}'$, where $\mathbf{C} = diag(e^{c_1}, \cdots, e^{c_p})$;

otherwise, let $\mathbf{\Sigma}_k^* = \mathbf{\Sigma}_{k-1}^*$ and $\mathbf{\Sigma}_k = \mathbf{\Sigma}_{k-1}$. Stop if $k+1$ is larger than a pre-set number M; otherwise replace k by $k+1$ and go to Step 1.

When the shrinkage prior is used to replace the constant prior for $\boldsymbol{\phi}$, the algorithms for Bayesian computation need to be modified by adding one step for drawing $\boldsymbol{\phi}$ using Fact 4. In cycle k, $\boldsymbol{\phi}_k$ is drawn in two-steps. First, parameter δ_k is drawn from an Inverse Gamma distribution, which depends on $\boldsymbol{\phi}_{k-1}$. Then $\boldsymbol{\phi}_k$ is drawn from a multivariate normal distribution that depends on δ_k, $\mathbf{\Sigma}_k$, and the data sample. The MCMC algorithm for numerical simulation of the posterior of

$(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ under the Minnesota-reference prior is based on the conditional posteriors in Fact 3 and Fact 5. The algorithm is quite similar to the algorithm used for drawing the posterior under the constant-reference prior combination, with a modification in the conditional density $\pi(\boldsymbol{\phi} \mid \boldsymbol{\Sigma}, \mathbf{Y})$.

5 MCMC Simulations

In the following we use numerical examples to evaluate the posteriors of competing estimators. We first generate $N = 1,000$ data samples from VARs with known parameters. Then for each generated data set we compute the Bayesian estimates under alternative priors via algorithms described in the previous section. The MCMC computations for eight prior combinations on $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ are labled as CA (Constant-RATS priors), CJ (Constant-Jeffreys priors), CR (Constant-Yang and Berger's Reference priors), SA (Shrinkage-RATS priors), SJ (Shrinkage-Jeffreys priors), SR (Shrinkage-Reference priors), TMR (Tight Minnesota-Reference priors), and LMR (Loose Minnesota-Reference priors).

The length of the Markov Chain is set at $M = 10,500$, with the first 500 cycles serving as burn-in runs. We choose a variety of data-generating VARs. Using the Monte Carlo results, we evaluate the Bayesian estimators under competing priors in terms of the frequentist risks, impulse responses, and the Mean Squared Errors of Forecast (MSEF). We also plot frequentist distributions of some elements of $\boldsymbol{\Sigma}$. We now discuss the criteria of evaluation in more detail.

5.1 Criteria for Evaluations of Bayesian VAR Estimates

a. Average Frequentist Losses. The most important criterion of evaluation is the frequentist risks of MLE and Bayesian estimators with various prior combinations on $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$. For loss function L_i, we denote the frequentist risk as $R_i (i = 1, 2)$. We also denote the estimates of $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ from the nth data set as $\widehat{\boldsymbol{\Sigma}}^n$ and $\widehat{\boldsymbol{\Phi}}^n$. The frequentist risks are estimated by averaging the losses over N samples:

$$R_1(\boldsymbol{\Sigma}) = \frac{1}{N} \sum_{n=1}^{N} L_1(\widehat{\boldsymbol{\Sigma}}^{(n)}, \boldsymbol{\Sigma}), \text{ and } R_2(\boldsymbol{\Phi}) = \frac{1}{N} \sum_{n=1}^{N} L_2(\widehat{\boldsymbol{\Phi}}^{(n)}, \boldsymbol{\Phi}).$$

b. Impulse Response Functions. A covariance stationary VAR has the moving average representation

$$\mathbf{y}_t = \mathbb{E}_0 \mathbf{y}_t + \sum_{j=0}^{t-1} \boldsymbol{\epsilon}_{t-j} \mathbf{H}_j, \tag{29}$$

where $\mathbf{H}_0$ is the p by p identity matrix, and the impulse responses of $\mathbf{y}_t$ to a shock $\boldsymbol{\epsilon}_{t-j}$ from j

periods earlier is

$$\mathbf{H}_j = \sum_{i=1}^{j} \mathbf{B}_i \mathbf{H}_{j-i}, \tag{30}$$

where $\mathbf{B}_i = 0$ for i larger than L. Note that the components of the vector of errors $\boldsymbol{\epsilon}_t$ are correlated since the covariance matrix is unrestricted. For example, the forecasting error of short-term interest rates may be correlated with that of inflation. Suppose a short-term interest rate is the monetary policy indicator. Then a monetary policy shock is represented by a shock in the short-term interest rate uncorrelated with other shocks. Thus more economic meaningful impulse responses are ones to orthogonalized (structural) errors. Orthogonalization of the errors can be achieved through the Cholesky decomposition of the covariance matrix,

$$\boldsymbol{\Sigma} = \boldsymbol{\Psi} \boldsymbol{\Psi}', \tag{31}$$

and a mapping from VAR errors to structural shocks,

$$\mathbf{u}_t = \boldsymbol{\epsilon}_t \boldsymbol{\Psi}^{-1'}. \tag{32}$$

The covariance matrix of the structural error vector $\mathbf{u}_t$ is the identity matrix. The impulse response to structural shocks from j periods earlier is given by $\mathbf{Z}_j = \boldsymbol{\Psi}' \mathbf{H}_j$. By definition, impulse responses are nonlinear functions of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$. The nonlinearity makes it difficult to derive frequentist inferences but does not pose difficulties for Bayesian simulations as long as $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ can be simulated. For the nth data set generated in the experiment, we denote the impulse response matrix on the ith step after the shock as $\widehat{\mathbf{Z}}_i^{(n)}$. The accuracy in estimation of the impulse responses (with forecasting horizon H) is measured by the frequentist average of sum of squared errors

$$R_{Imp} = \frac{1}{Np^2 H} \sum_{n=1}^{N} \mathrm{tr} \{ \sum_{i=1}^{H} (\mathbf{Z}_i - \widehat{\mathbf{Z}}_i^{(n)})'(\mathbf{Z}_i - \widehat{\mathbf{Z}}_i^{(n)}) \}. \tag{33}$$

c. Improvement in Mean Squared Errors of Forecast compared to the MLE. Besides risks, one frequentist criterion for evaluating estimators is the forecasting error attributable to the deviation of estimates from the true parameters. The h-step-ahead forecasting error at period T can be decomposed into two orthogonal parts:

$$\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h} \mid \widehat{\boldsymbol{\Phi}} = (\mathbf{y}_{T+h} - \widehat{\mathbf{y}}_{T+h} \mid \boldsymbol{\Phi}) + (\widehat{\mathbf{y}}_{T+h} \mid \boldsymbol{\Phi} - \widehat{\mathbf{y}}_{T+h} \mid \widehat{\boldsymbol{\Phi}}),$$

where $\widehat{\mathbf{y}}_{T+h} \mid \boldsymbol{\Phi}$ and $\widehat{\mathbf{y}}_{T+h} \mid \widehat{\boldsymbol{\Phi}}$ are the forecasts conditional on observations up to period T. They can be calculated from the VAR by setting the error term to zero after period T.

The first term in the right-hand-side is the sampling error. The second term is the forecasting error attributable to the deviation of estimates from the true parameters. Since the true parameters

are known, the second term can be calculated with competing estimators, and the MSEF of the second term can be compared. The MSEF is related to the frequentist loss in L_2. To see this, note the one-step-ahead forecasts are

$$\widehat{\mathbf{y}}_{T+1} \mid \boldsymbol{\Phi} \;=\; \mathbf{x}_T \boldsymbol{\Phi}, \tag{34}$$

$$\widehat{\mathbf{y}}_{T+1} \mid \widehat{\boldsymbol{\Phi}} \;=\; \mathbf{x}_T \widehat{\boldsymbol{\Phi}}. \tag{35}$$

Hence

$$\mathbb{E}(\widehat{\mathbf{y}}_{T+1} \mid \boldsymbol{\Phi} - \widehat{\mathbf{y}}_{T+1} \mid \widehat{\boldsymbol{\Phi}})'(\widehat{\mathbf{y}}_{T+1} \mid \boldsymbol{\Phi} - \widehat{\mathbf{y}}_{T+1} \mid \widehat{\boldsymbol{\Phi}}) = \mathbb{E}(\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}})' \mathbf{x}_T' \mathbf{x}_T (\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}). \tag{36}$$

In other words, the MSEF is the expectation of weighted quadratic estimation errors of $\boldsymbol{\Phi}$.

The frequentist average of the one-step-ahead MSEF for N samples is

$$\widehat{\mathbb{E}}(\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}})' \mathbf{x}_T' \mathbf{x}_T (\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}) = \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}^{(n)})' \mathbf{x}_T^{(n)'} \mathbf{x}_T^{(n)} (\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}^{(n)}).$$

5.2 Simulation Results

Numerous factors in the model design influence the performance of Bayesian estimators. For a given model, a larger sample size (T) makes smaller the effect of prior choice on the estimates. For a given sample size, a larger number of variables (p) included in the VAR or a longer lag length (L) makes the prior choice more important. Numerical results are presented to illustrate the effects of the sample size and dimension of the model. We will denote the VAR model (4) as $\mathrm{VAR}(T, p, L; \boldsymbol{\Phi}, \boldsymbol{\Sigma})$.

The relative performance of a prior also depends on the data generating process. For instance, the Minnesota prior should do well if data are generated from random walk processes and not as well otherwise. The types of models we choose have some characteristics commonly observed in macroeconomic time series. We first consider bivariate VARs with one lag. This setting involves the least number of parameters and allows for more experiments. We employ the covariance matrix $\boldsymbol{\Sigma}$ with different correlations and different types of VAR coefficient matrix $\boldsymbol{\Phi}$. We consider three types of data generating models for VARs with one lag: random walks with uncorrelated errors ($\boldsymbol{\Sigma} = \mathbf{I}_p$), Granger-causal chains with correlated errors, and VARs with relatively large off-diagonal elements in the lag coefficient matrix. In addition to the one-lag VARs, we also consider two-lag VARs that are close to being I(2) processes (i.e., the first difference in the time series are random walks).

Example 1 We consider $\mathrm{VAR}(T = 20, p = 2, L = 1; \boldsymbol{\Phi}, \boldsymbol{\Sigma})$, where $\boldsymbol{\Phi}$ is given by (3) with $\mathbf{c} = (0, 0), \mathbf{B}_1 = \mathbf{I}_2$, and $\boldsymbol{\Sigma} = \mathbf{I}_2$. This model serves as the benchmark. The assumption that the covariance matrix $\boldsymbol{\Sigma}$ is the identity matrix means that we treat the VAR disturbances as structural

shocks. The assumption that the VAR lag coefficient matrix is also the identity means that the VAR consists of independent random walk variables.

For 1,000 replications with Markov Chain length 10,500, the MCMC computations take about two hours total on a 1.7 GHz Pentium4 PC for all eight prior combinations. Simulation results are little changed when the Markov Chain length is reduced to 5,000 and the number of generated samples is reduced to 500, suggesting that the Markov chains converge rather quickly. The Metropolis-Hastings procedure is efficient for simulation of the $\boldsymbol{\Sigma}$ matrix under the Yang-Berger's reference prior, with acceptance rates around 58 percent.

The frequentist risks of MLE and Bayesian estimates under the test priors are reported in Table 1. The first two columns report the average and standard errors of losses associated with $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$ over the 1,000 generated samples. They show that the average losses associated with $\boldsymbol{\Sigma}$ are not influenced much by the prior on $\boldsymbol{\Phi}$. For estimating $\boldsymbol{\Sigma}$, the reference prior reduces risks by more than two third of that of the MLE and by about one half to two third compared with the Bayesian estimates under RATS and Jeffreys priors. For estimating $\boldsymbol{\Phi}$, the tight Minnesota prior does best and the loose Minnesota prior second best. This not surprising given the fact that the data generating $\boldsymbol{\Phi}$ is the mean of the Minnesota prior. If a researcher knows the data are generated by time series best characterized as random walks, then a tight Minnesota prior is a good choice. Comparison of the first four rows of the second column shows that Bayesian estimates of $\boldsymbol{\Phi}$ based on the constant prior are no better than the MLE. On the other hand, the shrinkage-reference prior reduces the average loss pertaining to $\boldsymbol{\Phi}$ by over a half, compared to the MLE. It is worthwhile to note that under the shrinkage prior the conditional posterior mean of $\boldsymbol{\phi}$ is $(\boldsymbol{\Sigma} \otimes (\delta \mathbf{X}'\mathbf{X})^{-1} + \mathbf{I}_J)^{-1}\widehat{\boldsymbol{\phi}}_{MLE}$, which appears to shrink $\widehat{\boldsymbol{\phi}}$ towards zero. However, not all elements of the shrinkage-based Bayesian estimator of the matrix $\boldsymbol{\Phi}$ are smaller than their MLE counterparts. It turns out that the diagonal elements of the Bayesian estimates of the lag coefficients $\mathbf{B}_1$ are larger than those of the MLE. In VAR models, by construction the regressors and the lags of error terms are correlated, and the MLEs are biased in finite samples. With a downward bias of the MLE of $\mathbf{B}_1$, the Bayesian estimator under the shrinkage prior may improve over the MLE by reducing bias for some elements and at the same time substantially reducing variances.

The third column of Table 1 reports the frequentist average of L_{22} losses associated with elements of the VAR lag coefficients $\mathbf{B}_1$. The difference between the second and third column is the average L_{21} loss associated with the constant terms in the VAR. The average losses in the third column are much smaller than the second column, suggesting that most of the L_2 losses are due to L_{21}. By definition, the intercept terms in the VAR do not affect the impulse responses. It is therefore reasonable that different rows of the fourth column of Table 1, which report the averages of mean

18

squared errors of elements of impulse responses, are fairly similar under different priors.

Figure 1 plots the impulse response of the first variable to a shock of the first equation. The response with respect to the true parameters is a horizontal line at 1. Each panel of the figure plots the frequentist average and frequentist standard errors of the impulse responses based on the MLE and Bayesian estimates. With the constant prior applied to $\boldsymbol{\Phi}$, the MLE and Bayesian estimates of $\boldsymbol{\Phi}$ are very similar. The impulse responses corresponding to the Bayesian estimates are different from those corresponding to the MLE mainly because the estimates of $\boldsymbol{\Sigma}$ are different. The MLEs of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ are both biased downwards. For this example the Bayesian estimate $\widehat{\boldsymbol{\Sigma}}_{CA}$ has a slight downward bias, $\widehat{\boldsymbol{\Sigma}}_{CJ}$ has an upward bias, and $\widehat{\boldsymbol{\Sigma}}_{CR}$ has almost no bias. The frequentist standard errors for the MLE appear to be slightly smaller than those of the Bayesian estimates largely because the average of $\widehat{\boldsymbol{\Sigma}}_{MLE}$ is significantly smaller than its Bayesian counterparts. In every panel there is a downward bias, which becomes worse with the increase of time. Overall the average errors in estimation of impulse responses are similar under the constant prior and MLE. Under the shrinkage prior the variance in $\boldsymbol{\Phi}$ is reduced. For similar estimates of $\boldsymbol{\Sigma}$, the impulse responses show smaller variances across samples. The improvements in the mean and reduced variance result in smaller frequentist average of the estimation errors in impulse responses. Such improvements are quite marginal, though, because the main source of estimation error is the downward bias of the MLE of $\boldsymbol{\Phi}$. The tight Minnesota prior is most effective in correcting the bias. Naturally the estimates of the tight Minnesota-reference prior yield the smallest average error in impulse responses.

Unlike the impulse responses, the forecast errors are affected by the estimates of the intercept terms. To illustrate the point, we will follow the notation in (3) and in addition let

$$\mathbf{y} = (\mathbf{y}_{T-1}, \cdots, \mathbf{y}_{T-L}), \ \mathbf{B} = \begin{pmatrix} \mathbf{B}_1 \\ \vdots \\ \mathbf{B}_L \end{pmatrix}, \ \Delta\mathbf{c} = \mathbf{c} - \widehat{\mathbf{c}}, \ \Delta\mathbf{B} = \mathbf{B} - \widehat{\mathbf{B}}. \tag{37}$$

The MSEF can be decomposed as

$$\mathbb{E}(\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}})' \mathbf{x}_T' \mathbf{x}_T (\boldsymbol{\Phi} - \widehat{\boldsymbol{\Phi}}) \ = \ \mathbb{E}(\Delta\mathbf{c}'\Delta\mathbf{c} + \Delta\mathbf{c}'\mathbf{y}\Delta\mathbf{B} + \Delta\mathbf{B}'\mathbf{y}'\Delta\mathbf{c} + \Delta\mathbf{B}'\mathbf{y}'\mathbf{y}\Delta\mathbf{B}). \tag{38}$$

As the frequentist average losses in Table 1 show, for the MLE and the constant-prior-based Bayesian estimators, the estimation error for the intercept term $\Delta\mathbf{c}$ is quite large compared to the error in the lag coefficients. This results in relatively large improvements of shrinkage-prior-based estimators over the MLE in forecasting errors, as indicated in the last column of Table 1.

Example 2 We now generate data sets from $\mathrm{VAR}(T = 20, p = 2, L = 1; \boldsymbol{\Phi}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0.71 \\ 0.71 & 2.0 \end{pmatrix}, \quad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & 1.0 \\ 0.7 & 0 \\ 0.3 & 1.0 \end{pmatrix}.$$

Here errors are assumed to have correlation of 0.5, and $\mathbf{B}_1$ is lower triangular, suggesting that the lags of $\mathbf{y}_{1t}$ are not useful for predicting $\mathbf{y}_{2t}$. The VAR contains a unit root.

We calculate the frequentist risks and compare the performance of Bayesian estimators based on the set of test priors. The results are reported in Table 2. Bayesian estimators based on the reference prior on $\mathbf{\Sigma}$ are better than the MLE, but the improvements are not as substantial as in the previous example. The less dominating performance of the reference prior may be due to the strong pairwise correlations of the VAR residuals in this example. By construction, the reference prior employed in this paper re-parameterizes $\mathbf{\Sigma}$ as $\mathbf{O}'\mathbf{D}\mathbf{O}$, with diagonal matrix $\mathbf{D}$ being the eigenvalues and $\mathbf{O}$ being an orthogonal matrix. The eigenvalues are placed before the orthogonal matrix in the order of importance, hence by design the performance for estimators for $\mathbf{D}$ is perceived to be more important. In the previous example the $\mathbf{\Sigma}$ matrix is diagonal, and the reference prior does much better. In this example, the pairwise correlations of VAR residuals are close to unity, hence the off-diagonal elements of the $\mathbf{\Sigma}$ matrix are more prominent. But note that even in this case the reference prior still does better than other priors. In Table 2 the relatively large forecasting errors of the estimates under the tight Minnesota-reference prior is caused by the poor estimates of the first column of $\mathbf{B}_1$, the column that corresponds to the variable that does not follow a random walk.

Example 3 We now consider $\mathrm{VAR}(T = 20, p = 2, L = 1; \mathbf{\Phi}, \mathbf{\Sigma})$, where

$$\mathbf{\Sigma} = \begin{pmatrix} 1.0 & 0.71 \\ 0.71 & 2.0 \end{pmatrix}, \text{ and } \mathbf{\Phi} = \begin{pmatrix} 1.0 & 1.0 \\ 0.3 & 0.7 \\ 0.7 & 0.3 \end{pmatrix}.$$

Here $\mathbf{B}_1$ is quite different from $\mathbf{I}_2$.

The focus is to compare the Minnesota prior with the shrinkage prior. The data-generating model in this example is substantially different from random walks. The tight Minnesota prior puts a heavy weight on the wrong prior information and results in rather poor estimates. The frequentist risks with respect to both $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ for the Bayesian estimates based on the tight Minnesota-reference prior are larger than those of the shrinkage-reference prior. The frequentist averages of the Bayesian estimates under the shrinkage-reference prior are

$$\widehat{\mathbf{\Sigma}}_{SR} = \begin{pmatrix} 1.1985 & 0.6904 \\ 0.6904 & 2.1723 \end{pmatrix}, \widehat{\mathbf{\Phi}}_{SR} = \begin{pmatrix} 0.6810 & 0.6252 \\ 0.3024 & 0.7006 \\ 0.7124 & 0.3159 \end{pmatrix}.$$

In comparison, the frequentist averages of the Minnesota-reference prior estimates deviate considerably further from the true parameters:

$$\widehat{\mathbf{\Sigma}}_{TMR} = \begin{pmatrix} 2.0674 & 0.0081 \\ 0.0081 & 2.8750 \end{pmatrix}, \widehat{\mathbf{\Phi}}_{TMR} = \begin{pmatrix} 0.6276 & 0.5878 \\ 0.8887 & 0.1458 \\ 0.1279 & 0.8714 \end{pmatrix}.$$

20

Under the tight Minnesota prior, the estimate of the VAR lag coefficient matrix is severely biased towards the identity matrix. The average estimation errors of impulse responses of the tight-Minnesota-prior-based estimates are larger than those of the other estimates. The last column of Table 3 shows that the one-step-ahead forecast errors of the Bayesian estimates under the tight Minnesota prior are much larger than those of the MLE.

The conditional densities $\pi(\phi \mid \Sigma, \mathbf{Y}, \delta)$ or $\pi(\phi \mid \Sigma, \mathbf{Y})$ under the shrinkage and Minnesota priors are both multivariate normal. Under the Minnesota prior, the conditional mean of ϕ, μ_M, is the MLE $\widehat{\phi}_{MLE}$ adjusted by the weighted difference between the mean of prior ϕ_0 and the MLE. Under the shrinkage prior the conditional mean μ_S is the MLE multiplied by a shrinkage matrix. Both the Minnesota prior and the shrinkage prior lead to smaller conditional variance. The reduction of conditional variance by the Minnesota prior depends on the variance of the prior M_0. The tighter the Minnesota prior (i.e. the smaller M_0) the larger is the reduction in the conditional variance. The relative performance of the shrinkage prior and the Minnesota prior depends on whether the Minnesota prior correctly reflects the true parameters. If ϕ_0 is closer to the true parameter ϕ than the MLE $\widehat{\phi}_{MLE}$, and if the variance M_0 is small, then the Minnesota prior should be superior to the shrinkage prior. On the other hand, if the Minnesota prior is not concentrated around the true parameter ϕ, then the shrinkage prior or the loose Minnesota prior may dominate the tight Minnesota prior.

Example 4   We generate data from VAR$(T = 20, p = 2, L = 1; \Phi, \Sigma)$, where

$$\Sigma = \begin{pmatrix} 1.0 & 0.71 \\ 0.71 & 2.0 \end{pmatrix}, \quad \Phi = \begin{pmatrix} 3.0 & 3.0 \\ 0.3 & 0 \\ 0 & 0.3 \end{pmatrix}.$$

Here $\mathbf{B}_1$ has small lag coefficients but large intercepts. In this case, the constant prior dominates the shrinkage prior and the Minnesota prior. The frequentist average of the MLE and Bayesian estimates exhibit a clear pattern. The averages of MLE over the 1000 samples are

$$\widehat{\Sigma}_{MLE} = \begin{pmatrix} 0.8553 & 0.6029 \\ 0.6029 & 1.7118 \end{pmatrix}, \quad \widehat{\Phi}_{MLE} = \begin{pmatrix} 3.4097 & 3.4608 \\ 0.1897 & -0.0021 \\ 0.0160 & 0.1952 \end{pmatrix}.$$

The estimates for Σ and the VAR lag coefficients are biased downwards. The intercepts are biased upwards. Under the constant-reference prior, the averages of the Bayesian estimates are

$$\widehat{\Sigma}_{CR} = \begin{pmatrix} 1.2143 & 0.6973 \\ 0.6973 & 2.2058 \end{pmatrix}, \quad \widehat{\Phi}_{CR} = \begin{pmatrix} 3.4093 & 3.4610 \\ 0.1898 & -0.0023 \\ 0.0160 & 0.1954 \end{pmatrix}.$$

The average of the estimates for $\boldsymbol{\Sigma}$ is better than that of the MLE, while the estimates for $\boldsymbol{\Phi}$ are almost identical to that of the MLE. Under the shrinkage-reference prior the averages of the estimates are

$$\widehat{\boldsymbol{\Sigma}}_{SR} = \begin{pmatrix} 1.3332 & 0.8558 \\ 0.8558 & 2.3866 \end{pmatrix}, \ \widehat{\boldsymbol{\Phi}}_{SR} = \begin{pmatrix} 1.9508 & 1.6142 \\ 0.4301 & 0.3080 \\ 0.0949 & 0.2892 \end{pmatrix}.$$

The estimates for $\boldsymbol{\Sigma}$ have an upward bias in magnitude similar to the downward bias of MLE. But the variance of the estimates is smaller than that of the MLE, which is the main reason for smaller frequentist risk associated with the Bayesian estimates. Contrary to the Bayesian estimates under the constant prior, under the shrinkage prior the Bayesian estimates for the intercepts are biased downward while the estimates for VAR coefficients tend to be biased upward. Finally, the frequentist average of the Bayesian estimates under the tight Minnesota-reference prior is

$$\widehat{\boldsymbol{\Sigma}}_{TMR} = \begin{pmatrix} 1.6204 & 0.9933 \\ 0.9933 & 2.9373 \end{pmatrix}, \ \widehat{\boldsymbol{\Phi}}_{TMR} = \begin{pmatrix} 0.6577 & 0.6273 \\ 0.8031 & 0.0539 \\ 0.0262 & 0.7732 \end{pmatrix}.$$

The estimates for the VAR lag coefficients are biased towards the identity matrix. The influence of the prior is significant because of the small sample size of the data.

In terms of the estimation errors of impulse responses and MSE of forecast, the constant prior also dominates the shrinkage and Minnesota priors, with the tight Minnesota prior being the worst among all priors under examination. The estimates of $\boldsymbol{\Sigma}$ and $\mathbf{B}_1$ both show upward bias under the shrinkage-reference prior, the compound effect of which may explain the relatively poor performance of the estimator in terms of impulse responses.

This example shows that for a VAR(1) (the number in the bracket indicates the lag length) model with large intercept terms and small VAR coefficients, the constant prior is better than the shrinkage or Minnesota prior. This result is partially due to fact that the downward biases of ML estimates of VAR lag coefficients are relatively small when the true parameters are near zero. MacKinnon and Smith (1998) show that the downward bias of ML estimates for an AR(1) coefficient is nonlinear in the true parameter. When the true parameter is near unity the downward bias is substantially larger than when the true parameter is near zero. The constant prior is better than the shrinkage and Minnesota priors in estimating the intercept terms. If the intercept terms are large, the downward bias induced by the shrinkage and Minnesota priors is amplified, resulting in undesirable performance. However, for most macroeconomic applications of VAR models, the first lag coefficient matrix $\mathbf{B}_1$ is not as small as in this example. So in practice, the dominance of the constant-prior is not a very likely scenario. In addition, the dominance of the constant prior is no

longer present for VARs with longer lags. For example, for the same covariance matrix and intercept terms, if the lag coefficient is changed from 0.3 in a VAR(1) to 0.1 in each of the lags in a VAR(3), then the shrinkage prior dominates the constant and Minnesota priors.

Example 5 We now generate data sets from $\mathrm{VAR}(T = 20, p = 2, L = 2; \boldsymbol{\Phi}, \boldsymbol{\Sigma})$, where

$$\boldsymbol{\Sigma} = \begin{pmatrix} 1.0 & 0 \\ 0 & 1.0 \end{pmatrix}, \qquad \boldsymbol{\Phi} = \begin{pmatrix} 1.0 & 1.0 \\ 1.85 & 0 \\ 0 & 1.85 \\ -0.9 & 0 \\ 0 & -0.9 \end{pmatrix}.$$

The VAR variables are nearly I(2). The averages of the estimates under the shrinkage-reference prior and the tight Minnesota-reference prior are

$$\widehat{\boldsymbol{\Sigma}}_{SR} = \begin{pmatrix} 1.0182 & 0.0380 \\ 0.0380 & 1.0029 \end{pmatrix}, \qquad \widehat{\boldsymbol{\Phi}}_{SR} = \begin{pmatrix} 0.6537 & 0.7291 \\ 1.5424 & 0.0365 \\ 0.0484 & 1.5321 \\ -0.6251 & 0.0099 \\ -0.0046 & -0.6194 \end{pmatrix}.$$

$$\widehat{\boldsymbol{\Sigma}}_{TMR} = \begin{pmatrix} 2.3509 & 0.1829 \\ 0.1829 & 2.2851 \end{pmatrix}, \qquad \widehat{\boldsymbol{\Phi}}_{TMR} = \begin{pmatrix} 0.2012 & 0.2727 \\ 1.1772 & 0.0245 \\ 0.0249 & 1.1719 \\ -0.2229 & 0.0026 \\ -0.0011 & -0.2229 \end{pmatrix}.$$

In this example the larger matrix $\boldsymbol{\Phi}$ does not substantially change the computation cost for the MCMC routine. The acceptance rates of the Metropolis step in simulation of $\boldsymbol{\Sigma}$ under the reference prior are around 63 percent. The tight Minnesota prior is again substantially worse than other priors in all aspects except for the average MSE of impulse responses. This is due to the fact that under the tight Minnesota prior the estimates for $\mathbf{B}_1$ are biased downward, while the estimates for $\boldsymbol{\Sigma}$ are biased upward. These two types of bias partially offset when the impulse response functions are computed. This example suggests that errors in estimating of impulse response functions may not be good indicators for accuracy of VAR estimates.

The five examples of the bivariate VAR provide a fairly comprehensive picture on the performance of the test priors. For estimating the covariance matrix $\boldsymbol{\Sigma}$, the reference prior dominates the Jeffreys and RATS priors. For estimating VAR coefficients $\boldsymbol{\Phi}$, the shrinkage prior most likely dominates the constant prior. The relative performance of the Minnesota prior depends on the tightness of the prior and the nature of the data generating models. When the data generating process is not similar to random walks, the tight Minnesota prior may be much less desirable than a loose Minnesota prior. In fact, when the data generating process is sufficiently different from the random walk, even the loose Minnesota prior can be undesirable (as in Example 4).

We examine the robustness of the pattern exhibited in Tables 1 to 5 by altering the sample size and the size of the VAR. We simulate the same models as that in Examples 1 to 5, but the sample size T is increased to 50 from 20. With the enlarged sample size, the average losses are smaller under all priors, and the difference in losses are smaller as well. This is because more data observations diminish the impact of prior choice. However, in most cases the shrinkage-reference prior still performs better than the other priors. We experiment with VARs containing more explosive roots and find that the shrinkage-reference prior combination still dominates other noninformative prior combinations.

The following examples show that the effects of prior choice are more prominent when the number of variables in the VAR is increased from two to six, even with sample size T increased from 20 to 50. We consider several VAR models representative of many monthly and quarterly macroeconomic variables. The first example is the same as Example 1, which consists of time series of random walks. We expect superior performance by the Minnesota prior since the prior centers at the true parameters. The second example combines the features of Examples 2 and 4. The third example is a VAR with two lags. In this example, we find that the shrinkage prior out-performs the Minnesota prior. Finally, we estimate a VAR using quarterly data of the U.S. economy and use the estimates as the "true" parameters to evaluate the fit of the estimators.

Example 6   We now consider VAR($T = 50, p = 6, L = 1; \boldsymbol{\Phi}, \boldsymbol{\Sigma}$), with intercept $\mathbf{c}=\mathbf{0}$, lag coefficients $\mathbf{B}_1=I_6$, and covariance matrix $\boldsymbol{\Sigma}=I_6$,

Compared to the case with $p = 2$ in Example 5, in this example there are a larger number of parameters. The number of parameters to be estimated in $\boldsymbol{\Sigma}$ is increased from 3 (with $p = 2$) to 21 (with $p = 6$) and the number of parameters in $\boldsymbol{\Phi}$ is increased from 6 to 42. The Bayesian estimators with the shrinkage-reference prior combination dominates MLE and Bayesian estimators under other priors in terms of average losses associated with the covariance matrix. The acceptance rates of the Metropolis step in simulating $\boldsymbol{\Sigma}$ under the reference prior are about 27 percent. Compared with Example 1, a notable difference made by the larger number of parameters and larger sample size is that the frequentist average loss associated with $\boldsymbol{\Phi}$ under the shrinkage-reference prior is now smaller than that of the MLE. It is known that MLE of $\mathbf{B}_1$ is biased towards the stationary region. The downward bias in $\mathbf{B}_1$ is much smaller under the shrinkage and Minnesota priors. Under the shrinkage prior, the frequentist average losses associated with $\boldsymbol{\Phi}$ are small mainly because the estimates of the intercept term are not as erratic as the MLE. A striking result is that the frequentist average loss for $\boldsymbol{\Phi}$ under the shrinkage-prior is smaller than that of the tight Minnesota prior. This is largely due to the fact that b_3, the variance of the Minnesota prior for the intercept term, is set at 1.0. If

b_3 is set at 0.2^2, then the average loss for $\mathbf{\Phi}$ is reduced from 2.653 to about 0.3, smaller than that under the shrinkage prior.

The frequentist risks of the Bayesian estimates of non-intercept terms under the shrinkage prior are larger than under the tight Minnesota prior and are comparable to those under the loose Minnesota prior. The tight Minnesota prior performs best in terms of impulse responses and forecasting errors. The shrinkage prior is effective in reducing the frequentist variance of the estimates, but it tends to yield biased estimates. The bias results in relatively mediocre performance in terms of impulse responses and forecast errors compared with the tight Minnesota prior.

A loose Minnesota prior, on the other hand, may not be better than the shrinkage prior. Table 6 shows that the average loss associated with the $\mathbf{\Phi}$ parameters under the loose Minnesota prior is 4.369, considerably better than the MLE average loss 15.681, but worse than the average loss of 1.258 under the shrinkage prior. Even though the Minnesota prior is centered close to the true parameters, if the prior is not tightly set, then the MLE estimates overwhelm the prior. As noted earlier, the small frequentist average losses of the $\mathbf{\Phi}$ parameters achieved by the shrinkage prior is through variance reduction for the intercept term. In this example, the shrinkage prior is not much inferior to the loose Minnesota prior for non-intercept terms. Overall, in comparison to the loosely specified Minnesota prior, the shrinkage prior is quite effective.

Table 6 demonstrates that the reference prior yields estimators for $\mathbf{\Sigma}$ with good frequentist properties in terms of average losses. More intuitive comparisons can be made by plotting the histograms of estimators of the $\mathbf{\Sigma}$ parameters across the 1,000 generated samples. Since it is impossible to plot such graphs for matrices, in the following we focus on a single element of covariance matrix $\mathbf{\Sigma}$, $\sigma_{1,1}$. Figure 2 plots the frequentist distributions of posterior means of $\sigma_{1,1}$ under test priors, and that of the MLE. Comparison of the panels shows that the MLE and the RATS-prior-based estimator are skewed to the left while the Jeffreys-prior-based estimators are more skewed to the right of the true value (1.0). The frequentist averages (standard errors) of estimates of $\sigma_{1,1}$ over the 1,000 samples are 0.755(0.168) for the MLE, 0.878 (0.195) for the Bayesian estimator with the shrinkage-RATS prior, 1.049 (0.233) for the estimator with shrinkage-Jeffreys prior, 0.936 (0.136) for the estimator with shrinkage-reference prior, and 0.921(0.130) for the estimator with the tight Minnesota-reference. The reference-prior-based estimator shows relatively small bias, but its most prominent feature is the small dispersion. The figure offers intuitive confirmations of results in Table 6. The figure indicates that the reference prior reduces average losses through variance reduction, not necessarily through bias reduction. Regarding the elements of $\mathbf{\Phi}$, it is well known that although shrinkage estimators may reduce risks, they do not improve universally over all elements. To provide a complete picture we should plot distributions of all elements of $\mathbf{\Phi}$, which will take up too much

space. Hence element-wise comparison of estimates for $\mathbf{\Phi}$ is omitted.

Example 7 We examine the test priors in a VAR with Granger causal chain. We consider VAR($T = 50, p = 6, L = 1; \mathbf{\Phi}, \mathbf{\Sigma}$), with intercept $\mathbf{c} = (1, 1, 1, 1, 1, 1)$, and following covariance matrix $\mathbf{\Sigma}$ and lag coefficients $\mathbf{B}_1$:

$$
\mathbf{\Sigma} = \begin{pmatrix}
1.00 & 0.71 & 0.87 & 1.00 & 1.12 & 1.22 \\
0.71 & 2.00 & 1.22 & 1.41 & 1.58 & 1.73 \\
0.87 & 1.22 & 3.00 & 1.73 & 1.94 & 2.12 \\
1.00 & 1.41 & 1.73 & 4.00 & 2.24 & 2.45 \\
1.12 & 1.58 & 1.94 & 2.24 & 5.00 & 2.74 \\
1.22 & 1.73 & 2.12 & 2.45 & 2.74 & 6.00
\end{pmatrix}, \ \mathbf{B}_1 = \begin{pmatrix}
1/6 & 0 & 0 & 0 & 0 & 0 \\
1/6 & 1/5 & 0 & 0 & 0 & 0 \\
1/6 & 1/5 & 1/4 & 0 & 0 & 0 \\
1/6 & 1/5 & 1/4 & 1/3 & 0 & 0 \\
1/6 & 1/5 & 1/4 & 1/3 & 1/2 & 0 \\
1/6 & 1/5 & 1/4 & 1/3 & 1/2 & 1
\end{pmatrix}.
$$

The covariance matrix implies pairwise correlation of 0.5. The VAR contains a unit root. The results are qualitatively the same as Example 2. The shrinkage prior produces better estimators of $\mathbf{B}_1$ than the MLE because it reduces variance through shrinkage. Many elements of the shrinkage-prior-based Bayesian estimator show smaller bias than the MLE. For example, the intercept terms in the MLE are considerably larger than those of the Bayesian estimators and larger than the true parameter of 1. The estimates under the shrinkage and Minnesota priors under-estimate the intercepts. The forecasts of the first variable by the shrinkage-prior-based estimators are worse than their constant-prior-based counterparts, similar to the finding in Example 4 where the constant prior is better when the VAR lag coefficients are small. It is not surprising that the estimator under the tight Minnesota prior does better in forecasting the sixth variable since it follows a random walk.

Example 8 We now consider VAR($T = 50, p = 6, L = 2; \mathbf{\Phi}, \mathbf{\Sigma}$), with intercept $\mathbf{c} = (1, 1, 1, 1, 1, 1)$, the covariance matrix $\mathbf{\Sigma}$ as in Example 7. The VAR lag coefficients $\mathbf{B}_1$ is twice the $\mathbf{B}_1$ matrix in Example 7, and $\mathbf{B}_2$ is the negative of the $\mathbf{B}_1$ matrix in Example 7. The sixth variable follows an I(2) process. For this example, we reduce the number of MCMC cycles to 5500 with 500 burn-in runs to reduce computing time (which is over eighty hours total for simulations under all priors). The acceptance rates for the Metropolis step in simulating $\mathbf{\Sigma}$ under the reference prior are about 36 percent. Table 8 shows that the Bayesian estimator of $\mathbf{\Phi}$ based on the tight Minnesota prior is better than the MLE, but for $\mathbf{B}_1$ it is worse than the MLE and the Bayesian estimator based on the shrinkage-reference prior. The loose Minnesota prior is better than the tight one in estimating the non-intercept terms of $\mathbf{\Phi}$ because the mean of the Minnesota prior is far from the true parameters. It is not surprising that it does better than the tight Minnesota prior in estimation errors of impulse responses. This is the opposite case of Example 6, in which a tight Minnesota prior is better when the prior is centered at the true parameters.

The examples show that the performance of the Minnesota priors depends on the data generating process and the setting of hyper-parameters. In practice, researchers often follow conventions when

26

they select hyper-parameter values. As we point out in the introduction, it is quite unlikely that a set of hyper-parameters is suitable for all data generating processes. The conventional values of the hyper-parameters (e.g., $b_1 = 0.2^2$) may result in undesirable estimators. On the other hand, when researchers decide to use alternative hyper-parameters to incorporate their knowledge of the data generating processes, it would become necessary for readers to take into account the difference between their own priors and those of the researchers. Adopting a noninformative prior as a reference for a wide range of empirical problems may be a better approach if a researcher is not very certain about the validity of his priors or when opinions of different researchers are diverse. In addition to the convenience in scientific reporting, a good noninformative prior may be less vulnerable to mistakes in researchers' judgement and therefore be able to deliver robust performance for a large variety of problems. The numerical examples show that the shrinkage prior produces more robust results compared to the Minnesota priors. The dominance of a tight Minnesota prior over the shrinkage prior in the random walk model is less remarkable than the dominance of the shrinkage prior over a "wrong" Minnesota prior.

Example 9   Now we consider a numerical example based on a set of actual macroeconomic data. We apply VAR$(T = 58, p = 6, L = 1; \mathbf{\Phi}, \mathbf{\Sigma})$ model to analyze quarterly data of the U.S. economy from 1987Q1 to 2001Q2. The variables include the M2 money stock, non-borrowed reserves, federal funds rate, world commodity price, GDP deflator, and real GDP. The commodity price data are obtained from the International Monetary Fund and the rest of data series from the FRED database at the Federal Reserve Bank of St. Louis. All variables except the fed funds rate are growth rates. All variables are measured in percentage terms. These variables frequently appear in macroeconomics related VARs (e.g. Sims 1992, Gordon and Leeper 1994, Sims and Zha 1998b, and Christiano, Eichenbaum, and Evans 1999). The six data series exhibit strong pairwise and serial correlations. We use the MLE of the actual data as the "true" parameters for $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ and conduct the same MCMC simulations for drawing posteriors of VAR coefficients and the covariance matrix as in previous examples. Note that the impulse responses are based on the lower triangular mapping from the VAR residuals to structural shocks. The order of the variables implies that a shock in a variable affects all other variables placed before it contemporaneously but not the other way around.

The reference prior shows moderate improvement over the RATS prior and is comparable to the Jeffreys prior. The absence of more significant improvement of the reference prior can be explained by two reasons. First, there are strong pairwise correlations of the VAR error terms that make the off-diagonal elements prominent. Since the reference prior places the variance components in higher priority than the covariance components, it tends to perform less well in case the covariance components are large. Second, the reference prior shrinks the eigenvalues of the covariance towards

one another. It does less well when the true data generating model has variance components that are very different in scale, as is the case here. The variances of the error terms range from .035 (GDP deflator) to 7.74 (commodity price).

The Bayesian estimators of $\boldsymbol{\Phi}$ show significant improvement over the MLE. The shrinkage prior does slightly better than the Minnesota priors for the VAR lag coefficients. In this example, the VAR has one lag, and the "true" parameters do not deviate too much from random walks. We also estimate a VAR model with two lags using the same data set. If the VAR(2) estimates are used as the "true" parameters then the mean of the Minnesota prior significantly deviates from these true parameters. In this case, the shrinkage prior is much better than the Minnesota priors.

A few general conclusions can be drawn from these numerical examples. (1) Yang and Berger's reference prior for the covariance matrix $\boldsymbol{\Sigma}$ dominates the Jeffreys and RATS prior in many cases. The reference prior does less well when the data-generating $\boldsymbol{\Sigma}$ has large off-diagonal elements and the variance components are significantly different. But even in the least favorable cases, the reference prior is not dominated by its competitors. (2) The posterior mean of $\boldsymbol{\Phi}$ under the constant prior (regardless of the prior on $\boldsymbol{\Sigma}$) has properties very similar to the MLE. For VAR(1) models consisting near-random-walk type variables, the frequentist averages of the posterior means under the constant prior over-estimate the intercept term $\mathbf{c}$ and under-estimate the VAR lag coefficients $\mathbf{B}_1$. The frequentist averages of quadratic losses are large for these estimators because the variances of the intercept terms across samples are often quite large. The shrinkage prior, on the other hand, tends to under-estimate the intercept terms and over-estimate the VAR lag coefficients. The shrinkage-prior-based estimators induce smaller frequentist average losses mainly because the shrinkage prior effectively reduces variances of the elements in $\boldsymbol{\Phi}$ across samples. (3) Impulse responses and forecasting errors are nonlinear functions of elements of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$. Smaller frequentist average losses with respect to parameters do not necessarily lead to smaller average losses in terms of impulse responses and forecasting errors, and vice versa. In Example 5, the tight Minnesota prior happens to significantly over-estimate $\boldsymbol{\Sigma}$ and under-estimate $\mathbf{B}_1$. But the biases cancel out and the estimates for impulse responses are more accurate than those with better estimated $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$. A shrinkage prior often reduces the variance of the elements of the posterior mean of $\boldsymbol{\Phi}$ but may make them quite biased. The bias may result in poor performance in terms of impulse responses and forecasting errors. Estimators other than the posterior mean may be more desirable under the shrinkage prior if they can reduce the bias. (4) As with any informative prior, the performance of the Minnesota prior depends on the nature of the data generating model and the hyper-parameters. If the VAR is made of random-walk type of variables, then a tightly set Minnesota prior does better than a loosely set Minnesota prior and noninformative priors. However, if the model is not in agreement

with the prior, a tightly set Minnesota prior does much worse than alternative priors. The examples highlight the sensitivity of the estimates to the hyper-parameters and serve as a note of caution for researchers who rely on an informative prior.

6   Concluding Remarks

In this study we evaluate Bayesian VAR estimators based on several noninformative priors in terms of frequentist risks. For the VAR covariance matrix $\mathbf{\Sigma}$, we study the Jeffreys prior, the RATS prior and Yang and Berger's reference prior. For VAR coefficients $\mathbf{\Phi}$, we consider the constant prior, a shrinkage prior, and the Minnesota prior. We establish the propriety of posteriors as well as existence of posterior moments for $(\mathbf{\Phi}, \mathbf{\Sigma})$ under a general class of priors that includes the prior combinations studied in this paper. We compute posteriors under different priors via MCMC simulations. Our numerical examples show that in most cases the combination with the shrinkage prior on $\mathbf{\Phi}$ and Yang and Berger's reference prior on $\mathbf{\Sigma}$ produces smaller frequentist average losses than other combinations of noninformative priors, mainly through reducing the variances of estimates across samples. In all examples considered in the paper the constant prior generates Bayesian estimates of $\mathbf{\Phi}$ very similar to the MLE. We also find that the performance of the Minnesota prior critically depends on the tightness of the prior and the nature of data generating models. A tightly set Minnesota prior dominates the shrinkage prior when the data generation processes are close to random walks, while the shrinkage prior or a loosely set Minnesota prior is a better choice otherwise. We have argued in the introduction that Bayesian procedures with appropriate priors are a practical tool for users of VAR models who are mainly concerned with finite sample properties of estimators. In light of the MCMC simulation results, we conclude that the shrinkage-reference prior combination is a reasonable choice for Bayesian analysis of finite sample inferences of VAR models.

The present study can be extended in several directions. First, it is useful to explore other priors for the VAR model. For estimation of identified VARs, identifying restrictions on the factorization of the covariance matrix $\mathbf{\Sigma}$ may be incorporated into a prior in a way similar to Sims and Zha (1998a, 1999). For the VAR coefficients $\mathbf{\Phi}$, it is useful to investigate whether the shrinkage prior can be modified for better bias correction. Note that the present paper considers priors for $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ separately. Joint noninformative priors for $(\mathbf{\Phi}, \mathbf{\Sigma})$ are more difficult to derive. Consider the AR(1) model $y_t = \beta y_{t-1} + \epsilon_t$, where ϵ_t is iid normal with variance σ^2. The asymptotic form of Berger-Bernardo's reference prior for (β, σ) is $(1 - \beta^2)^{-1/2}\sigma^{-1}$ in the stationary region $|\beta| < 1$, which takes the same form as the Jeffreys prior. Jeffreys (1967) deems the performance of his prior in multiparameter cases unsatisfactory. The Jeffreys and reference prior in this model put infinite

weight at the unit root. Zellner's (1997) MDIP takes the more reasonable form of $(1 - \beta^2)^{1/2}\sigma^{-1}$. For the finite sample AR(1) model, Phillips (1991) derives the joint Jeffreys prior, and Berger and Yang (1996) derive a joint reference prior for the autoregressive and the variance parameters. Sims (1991) points out some undesirable features of the finite sample AR(1) Jeffreys prior. Nonetheless, deriving and evaluating joint priors for the VAR model is an interesting research topic.

The second direction of extension is to consider loss functions that produce Bayesian estimators different from the posterior mean. There are good reasons to doubt the use of the constant-weighted quadratic loss. In economic applications, the elements in matrix $\mathbf{\Phi}$ are unlikely to be of equal importance. Furthermore, if the unit of measurement is changed for a data series (e.g., the dollar amount of GDP is measured in trillions instead of billions), then the corresponding elements in $\mathbf{\Phi}$ also change in magnitude. It is obvious that placing data-independent weights on the estimation errors is unreasonable. Some alternatives to the quadratic loss function include Zellner's (1986) LINEX asymmetric loss and functions used for the Minimum Expected Loss (MELO) approach in Zellner (1978). The LINEX loss allows for asymmetric weight on the positive and negative estimation errors, and the MELO functions place data-dependent weights on the elements of $\mathbf{\Phi}$. An additional motivation for considering alternative loss functions is that the posterior mean of $\mathbf{\Phi}$ under the shrinkage prior can be quite biased. Correction of the bias may make substantial improvement for estimation of the impulse responses. These questions are beyond the scope of this paper, and they are on our agenda for future research.

References

Anderson, T.W. (1984). *An Introduction to Multivariate Statistical Analysis.* 2nd edition, Wiley, New York.

Anderson, T.W., Olkin, I. and Underhill, L.G. (1987). Generation of random orthogonal matrices, *SIAM J. Sci. Statist. Comput.*, **8** 625-629.

Baranchik, A. J. (1964). Multiple regression and estimation of the mean of multivariate normal distribution. Technical Report 51, Dept. Statistics, Stanford University.

Berger, J.O. (1984). *Statistical Decision Theory and Bayesian Analysis.* 2nd Edition. Springer-Verlag, New York.

Berger, J.O. and Bernardo, J.M. (1989). Estimating a product of means: Bayesian analysis with reference priors. *Journal of the American Statistical Association*, **84**, 200-207.

Berger, J.O. and Bernardo, J.M. (1992). On the development of reference priors. In *Bayesian Analysis IV*, J.M. Bernardo, et. al., (Eds.). Oxford University Press, Oxford.

Berger, J.O. and Strawderman, W.E. (1996). Choice of hierarchical priors: Admissibility in estimation of normal means. *Annals of Statistics,* **24,** 931-951.

Berger, J.O. and Yang, R. (1996). Noninformative priors and Bayesian testing for the AR(1) model. *Econometric Theory* **10**, 461-482.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *J. Roy. Statist. Soc. Ser. B*, **41**, 113-147.

Chib, S. (1998). Estimation and comparison of multiple change point models. *Journal of Econometrics*, **86**, 221-241.

Chib, S. and Hamilton, B. (2000). Bayesian analysis of cross section and clustered data treatment models. *Journal of Econometrics*, **97**, 25-50.

Deschamps, P. J. (2000). Exact small-sample inference in stationary, fully regular, dynamic models. *Journal of Econometrics* **97**, 51-91.

Geisser, S. (1965). Bayesian estimation in multivariate analysis. *Annals of Mathematical Statistics* **36**, 150-159.

Gelfand, A.E. and Smith, A.F.M. (1990). Sampling based approaches to calculating marginal densities. *Journal of the American Statistical Association* **85**, 398-409.

Geweke, J. (1996). Monte Carlo Simulation and Numerical Integration. In Amman, H. M., Kendrick, D. A., Rust, J., eds. *Handbook of computational economics.* **Volume 1**. Elsevier Science, North-Holland, Amsterdam, New York and Oxford. 731-800.

Geweke, J. (1999). Using simulation methods for Bayesian econometric models: inference, development, and communication. *Econometric Reviews*, **18**, 1-73.

Gordon, D. B. and Leeper, E. M. (1994), The dynamic impacts of monetary policy: an exercise in tentative identification, *Journal of Political Economy* **102**, 1228–1247.

Hamilton, J. D. (1994) *Time Series Analysis.* Princeton University Press, Princeton, New Jersey.

Hobert, J.P. and Casella, G. (1996). The effect of improper priors on Gibbs sampling in hierarchical linear mixed models. *Journal of the American Statistical Association* **91**, 1461-1473.

Jeffreys, H. (1967) *Theory of Probability.* Oxford University Press, London.

Kadiyala, K.R. and Karlsson, S. (1997). Numerical methods for estimation and inference in Bayesian VAR-models. *Journal of Applied Econometrics.* **12,** 99-132.

Kass, R.E. and Wasserman, L. (1996). The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, **91**, 1343-1370.

Kilian, L. (1999). Finite sample properties of percentile and percentile-t bootstrap confidence intervals for impulse responses. *Review of Economics and Statistics*, **81**, 652-660.

Lee, K. and Ni, S. (2002). On the dynamic effects of oil price shocks–a study using industry level data. *Journal of Monetary Economics,* **49**, 823-852.

Leeper, E. M. and Zha T. (1999). Modest policy interventions. *Federal Reserve Bank of Atlanta working paper* **99-22**.

Litterman, R.B. (1986). Forecasting with Bayesian vector autoregression-five years of experience. *Journal of Business and Economic Statistics*, **4**, 25-38.

MacKinnon, J.G. and Smith, A. A. Jr. (1998). Approximate bias correction in econometrics, *Journal of Econometrics* **85**, 205– 230.

Neyman, J. and Scott, E.L. (1948). Consistent estimates based on partially consistent observations. *Econometrica* **16,** 1-32.

Pagan, A. R. and Robertson, J. C. (1998). Structural models of the liquidity effect, *The Review of Economics and Statistics* **80**, 202 – 217.

Phillips, P.C.B. (1991). To criticize the critics: an objective Bayesian analysis of stochastic trends. *Journal of Applied Econometrics*, **6**, 423-434.

Press, J.S. (1982). *Applied multivariate analysis: Using Bayesian and Frequentist methods of inference (Second edition)* Krieger, Florida.

Sims, C.A. (1972). Money, income, and causality. *The American Economic Review* **62,** 540-552.

Sims, C.A. (1980). Macroeconomics and reality, *Econometrica*, **48**, 1-48.

Sims, C.A. (1986). Are forecast models usable for policy analysis? *Quarterly Review of Federal Reserve Bank of Minneapolis.*

Sims, C.A. (1991). Comment by Christopher A. Sims on 'To criticize the critics'. by Peter C.B. Phillips. *Journal of Applied Econometrics*, **6**, 423-434.

Sims, C.A. and Zha T. (1998a). Bayesian methods for dynamic multivariate models, *International Economic Review*, **39**, 949-968.

Sims, C.A. and Zha T. (1998b). Does monetary policy generate recessions? *Federal Reserve Bank of Atlanta working paper* **98-12**.

Sims, C. A. and Zha T. (1999). Error Bands for Impulse Responses, *Econometrica*, **67**, 1113–1155.

Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution, *Proceedings of the third Berkeley symposium*, **Vol. 1**, 197-206, Berkeley: University of California Press.

Sun, D. and Berger, J.O. (1998). Reference priors under partial information. *Biometrika*, **85**, 55-71.

Sun, D. and Ye, K. (1995). Reference prior Bayesian analysis for normal mean products. *Journal of the American Statistical Association,* **90**. 589-597.

Sun, D., Tsutakawa, R.K. and He, H. (2001). Propriety of posteriors with improper priors in hierarchical linear mixed models. *Statistica Sinica,* **11**, 77-95.

Sun, D. and Ni, S. (2001). Bayesian analysis of vector-autoregressive models with noninformative priors. Submitted.

Tiao, G.C. and Zellner, A. (1964). On the Bayesian estimation analysis of multivariate regression, *Journal of Royal Statistical Society, B,* **26,** 389-399.

Yang, R. and Berger, J.O. (1994). Estimation of a covariance matrix using the reference prior. *Annals of Statistics,* **22,** 1195-1211.

Zellner, A (1971). *An Introduction to Bayesian Inference in Econometrics.* John Wiley & Sons, New York.

Zellner, A (1978). Estimation of functions of population means and regression coefficients including structural coefficients: a minimum expected loss approach. *Journal of Econometrics,* **8,** 127-158.

Zellner, A (1986). Bayesian estimation and prediction using asymmetric loss functions. *Journal of the American Statistical Association,* **81**. 446-451.

Zellner, A (1997) Maximal data information prior distributions, Chapter 8 of *Bayesian Analysis in Econometrics and Statistics,* Edward Elgar, Lyme, UK.

Appendix A: Proof for Theorem 1

In the following, we let $C_1, C_2, \ldots$ be constants depending only on sample size T and observation $\mathbf{Y}$. We rewrite the likelihood function (5) of $(\boldsymbol{\phi}, \boldsymbol{\Sigma})$ as

$$L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \;\; = \;\; \frac{1}{|\boldsymbol{\Sigma}|^{T/2}} \exp\left[-\frac{1}{2}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}})'\left\{\boldsymbol{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})\right\}(\boldsymbol{\phi} - \widehat{\boldsymbol{\phi}}) - \frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\}\right], \qquad (39)$$

where $\widehat{\boldsymbol{\phi}} = vec(\widehat{\boldsymbol{\Phi}})$ is the ML estimator, where $\widehat{\boldsymbol{\Phi}}$ and $\mathbf{S}(\widehat{\boldsymbol{\Phi}}) = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Phi}})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Phi}})$ are given by (6) and (7) respectively. Then

$$\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})d\boldsymbol{\phi} \;\; = \;\; \frac{(2\pi)^{J/2}}{|\boldsymbol{\Sigma}|^{T/2}|\boldsymbol{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})|^{1/2}} \, \mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\right\}.$$

Since $|\boldsymbol{\Sigma}^{-1} \otimes (\mathbf{X}'\mathbf{X})| = |\boldsymbol{\Sigma}|^{-(Lp+1)}|\mathbf{X}'\mathbf{X}|^p$,

$$\int\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_{(0,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma})d\boldsymbol{\phi}\,d\boldsymbol{\Sigma} \;\; \leq \;\; C_1 \int \frac{\mathrm{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\right\}}{|\boldsymbol{\Sigma}|^{(T-Lp-1+b)/2}\{\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j)\}^c}d\boldsymbol{\Sigma}. \quad (40)$$

Use the orthogonal decomposition $\boldsymbol{\Sigma} = \mathbf{O}'\boldsymbol{\Lambda}\mathbf{O}$, where $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$, and $\mathbf{O}$ is an orthogonal matrix of the form $\mathbf{O} = (\mathbf{O}_{12}\mathbf{O}_{13}\cdots\mathbf{O}_{1p})(\mathbf{O}_{23}\cdots\mathbf{O}_{2p})\cdots(\mathbf{O}_{p-1,p})$. Each $\mathbf{O}_{ij}$ is a simple orthogonal matrix of the form

$$\mathbf{O}_{ij} = \mathbf{O}_{ij}(o_{ij}) \;\; = \;\; \begin{array}{c} \\ i \\ \\ j \\ \\ \end{array}\begin{pmatrix} \mathbf{I} & 0 & 0 & 0 & 0 \\ 0 & \cos(o_{ij}) & 0 & -\sin(o_{ij}) & 0 \\ 0 & 0 & \mathbf{I} & 0 & 0 \\ 0 & \sin(o_{ij}) & 0 & \cos(o_{ij}) & 0 \\ 0 & 0 & 0 & 0 & \mathbf{I} \end{pmatrix}, \\ \qquad\quad i \qquad\qquad j$$

where $o_{ij} \in [-\pi/2, \pi/2]$. Let $\boldsymbol{\lambda} = (\lambda_1, \cdots, \lambda_p)$ and $\mathbf{o} = (o_{ij}, 1 \leq i < j \leq p)$. It follows from Anderson,

Olkin and Unherhill (1987) that the transformation from $\boldsymbol{\Sigma}$ to $(\boldsymbol{\lambda}, \mathbf{o})$ has the Jacobian

$$|\mathbf{J}| \equiv \left\{ \prod_{1 \leq i < j \leq p} \cos^{j-i-1}(o_{ij}) \right\} \left\{ \prod_{1 \leq i < j \leq p} (\lambda_i - \lambda_j) \right\}. \tag{41}$$

So the right hand side of (40) equals

$$C_1 \int \int \left\{ \prod_{1 \leq i < j \leq p} \cos^{j-i-1}(o_{ij}) \right\} \frac{\{\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j)\}^{1-c}}{\prod_{i=1}^{p} \lambda_i^{(T-Lp-1+b)/2}} \operatorname{etr}\left\{ -\frac{1}{2} \boldsymbol{\Lambda}^{-1} \mathbf{O} \mathbf{S}(\widehat{\boldsymbol{\Phi}}) \mathbf{O}' \right\} d\boldsymbol{\lambda} d\mathbf{o}$$

$$\leq C_1 \int \frac{\{\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j)\}^{1-c}}{\prod_{i=1}^{p} \lambda_i^{(T-Lp-1+b)/2}} \operatorname{etr}\left\{ -\frac{1}{2} \boldsymbol{\Lambda}^{-1} \mathbf{O} \mathbf{S}(\widehat{\boldsymbol{\Phi}}) \mathbf{O}' \right\} d\boldsymbol{\lambda} d\mathbf{o}. \tag{42}$$

The last inequality holds because $|\cos^{j-i-1}(o_{ij})| \leq 1$.

Let $\eta_1 > \eta_2 > \cdots > \eta_p > 0$ be the eigenvalues of $\mathbf{S}(\widehat{\boldsymbol{\Phi}})$, so that $\mathbf{S}(\widehat{\boldsymbol{\Phi}}) = \boldsymbol{\Gamma} \operatorname{diag}(\eta_1, \eta_2, \cdots, \eta_p) \boldsymbol{\Gamma}'$, where $\boldsymbol{\Gamma}$ is a $p \times p$ orthogonal matrix. Clearly $\mathbf{S}(\widehat{\boldsymbol{\Phi}}) - \eta_p \mathbf{I}_p$ is nonnegitive definite, and

$$\operatorname{tr}(\boldsymbol{\Lambda}^{-1} \mathbf{O} \mathbf{S}(\widehat{\boldsymbol{\Phi}}) \mathbf{O}') \geq \operatorname{tr}(\boldsymbol{\Lambda}^{-1} \mathbf{O} \eta_p \mathbf{I}_p \mathbf{O}') = \eta_p \operatorname{tr}(\boldsymbol{\Lambda}^{-1}) = \sum_{j=1}^{p} \frac{\eta_p}{\lambda_j}. \tag{43}$$

Combining (40), (42) and (43), we have

$$\int \int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \pi_{(0,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) d\boldsymbol{\phi} \, d\boldsymbol{\Sigma} \leq C_2 \int \frac{\left\{ \prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j) \right\}^{1-c}}{\prod_{i=1}^{p} \lambda_i^{(T-Lp-1+b)/2}} \exp\left(-\sum_{j=1}^{p} \frac{\eta_p}{2\lambda_j} \right) d\boldsymbol{\lambda} d\mathbf{o}$$

$$\leq C_3 \int \frac{\left\{ \prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j) \right\}^{1-c}}{\prod_{i=1}^{p} \lambda_i^{(T-Lp-1+b)/2}} \exp\left(-\sum_{j=1}^{p} \frac{\eta_p}{2\lambda_j} \right) d\boldsymbol{\lambda}. \tag{44}$$

The last inequality holds because the range of o_{ij} is bounded.

If $c = 0$, note that $\prod_{1 \leq i < j \leq p}(\lambda_i - \lambda_j) \leq \prod_{i=1}^{p} \lambda_i^{p-i}$, and the right hand side of (44) is bounded above by

$$C_3 \int \left\{ \prod_{i=1}^{p} \lambda_i^{p-i} \right\} \prod_{i=1}^{p} \frac{1}{\lambda_i^{(T-Lp-1+b)/2}} \exp\left(-\sum_{j=1}^{p} \frac{\eta_p}{2\lambda_j} \right) d\boldsymbol{\lambda}$$

$$= C_3 \prod_{i=1}^{p} \int_0^{\infty} \frac{1}{\lambda_i^{(T-Lp-1+b-2p+2i)/2}} \exp\left(-\frac{\eta_p}{2\lambda_i} \right) d\lambda_i. \tag{45}$$

Note that $\int_0^{\infty} x^{-(\alpha+1)} x^{-\beta/x} dx$ is finite if and only if $\alpha > 0$ and $\beta > 0$. So the right hand side is integrable if $T - Lp - 1 + b - 2p + 2 > 2$, which holds if $T > (L+2)p + 1 - b$.

If $c = 1$, (44) becomes

$$\int \int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \pi_{(0,b,1)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) d\boldsymbol{\phi} \, d\boldsymbol{\Sigma} \leq C_3 \prod_{i=1}^{p} \int_0^{\infty} \frac{1}{\lambda_i^{(T-Lp-1+b)/2}} \exp\left(-\frac{\eta_p}{2\lambda_i} \right) d\lambda_i, \tag{46}$$

which is integrable if $T - Lp - 1 + b - 2 > 0$, i.e. $T > Lp + 3 - b$. The results then follow.

34

Appendix B: Proof for Theorem 3

Using the expression (39) of the likihood function and the hierarchical structure of (10), we have

$$
\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_{(a)}(\boldsymbol{\phi})\, d\boldsymbol{\phi} = \int_0^\infty \left\{ \int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_s(\boldsymbol{\phi}\,|\,\delta)\, d\boldsymbol{\phi} \right\} \pi_a(\delta) d\delta
$$

$$
= \int_0^\infty \frac{(2\pi)^{J/2}|\boldsymbol{\Sigma}|^{-T/2}}{\delta^{J/2}|\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X}) + \delta^{-1}\mathbf{I}_J|^{1/2}} \operatorname{etr}\left\{ -\frac{\widehat{\boldsymbol{\phi}}'\mathbf{G}\widehat{\boldsymbol{\phi}}}{2} - \frac{\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})}{2} \right\} \pi_a(\delta) d\delta,
$$

where

$$
\begin{aligned}
\mathbf{G} &= \boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X}) - \{\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X})\}\{\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X}) + \delta^{-1}\mathbf{I}_J\}^{-1}\{\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X})\} \\
&= \delta^{-1}\{\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X}) + \delta^{-1}\mathbf{I}_J\}^{-1}\{\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X})\} \\
&= \{\delta\mathbf{I}_J + \boldsymbol{\Sigma}\otimes(\mathbf{X}'\mathbf{X})\}^{-1}.
\end{aligned} \tag{47}
$$

Clearly, $\mathbf{G}$ is nonnegative definite and $\operatorname{etr}\{-\frac{1}{2}\widehat{\boldsymbol{\phi}}'\mathbf{G}\widehat{\boldsymbol{\phi}}\} \leq 1$. Define $\boldsymbol{\Lambda} = diag(\lambda_1, \ldots, \lambda_p)$ and $\boldsymbol{\Xi} = diag(\xi_1, \cdots, \xi_{Lp+1})$, where $\lambda_1 > \cdots > \lambda_p$ are the eigenvalues of $\boldsymbol{\Sigma}$ and $\xi_1 \geq \cdots \geq \xi_{Lp+1} > 0$ are the eigenvalues of the matrix $\mathbf{X}'\mathbf{X}$. Then

$$
\begin{aligned}
\delta^{J/2}|\boldsymbol{\Sigma}^{-1}\otimes(\mathbf{X}'\mathbf{X}) + \delta^{-1}\mathbf{I}_J|^{1/2} &= |\delta\boldsymbol{\Lambda}^{-1}\otimes\boldsymbol{\Xi} + \mathbf{I}_J|^{1/2} = \prod_{i=1}^{p}\prod_{j=1}^{Lp+1}(\delta\xi_j\lambda_i^{-1} + 1)^{1/2} \\
&\geq \prod_{i=1}^{p}(\delta\xi_{Lp+1}\lambda_i^{-1} + 1)^{(Lp+1)/2} \\
&\geq (\delta\xi_{Lp+1}\lambda_p^{-1} + 1)^{J/2}.
\end{aligned}
$$

So we have

$$
\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_{(a)}(\boldsymbol{\phi})d\boldsymbol{\phi} \leq \frac{1}{(2\pi)^{J/2}|\boldsymbol{\Sigma}|^{T/2}} \operatorname{etr}\left\{ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}}) \right\} \int_0^\infty \frac{\delta^{(J-2-a)/2}}{(\delta\xi_{Lp+1}\lambda_p^{-1} + 1)^{J/2}} d\delta.
$$

Making the transformation $u = \delta\xi_{Lp+1}\lambda_p^{-1}/(\delta\xi_{Lp+1}\lambda_p^{-1} + 1)$, we get $\delta = (\lambda_p/\xi_{Lp+1})u/(1-u)$. Thus

$$
\begin{aligned}
\int_0^\infty \frac{\delta^{(J-2-a)/2}}{(\delta\xi_{Lp+1}\lambda_p^{-1} + 1)^{J/2}} d\delta &= \left(\frac{\lambda_p}{\xi_{Lp+1}}\right)^{(J-a)/2} \int_0^1 \left(\frac{u}{1-u}\right)^{(J-2-a)/2}(1-u)^{J/2}du \\
&= \left(\frac{\lambda_p}{\xi_{Lp+1}}\right)^{(J-a)/2} \int_0^1 u^{(J-a)/2-1}(1-u)^{a/2+1}du \\
&= \left(\frac{\lambda_p}{\xi_{Lp+1}}\right)^{(J-a)/2} Beta\left(\frac{J-a}{2}, \frac{a}{2}+2\right).
\end{aligned}
$$

The last equality holds from Condition (A). So

$$
\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_{(a)}(\boldsymbol{\phi})d\boldsymbol{\phi} \leq C\frac{\lambda_p^{(J-a)/2}}{|\boldsymbol{\Sigma}|^{T/2}} \operatorname{etr}\left\{ -\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}}) \right\}, \tag{48}
$$

where $C = Beta(\frac{1}{2}(J-a), \frac{1}{2}a+2)/\{(2\pi)^{J/2}\xi_{Lp+1}^{(J-a)/2}\}$. Since $\pi_{(a,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) = \pi_{(a)}(\boldsymbol{\phi})\pi_{(b,c)}(\boldsymbol{\Sigma})$, we have

$$
\int\int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma})\pi_{(a,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma})d\boldsymbol{\phi}\, d\boldsymbol{\Sigma}
$$

$$\leq \quad C \int \frac{\lambda_p^{(J-a)/2}}{|\mathbf{\Sigma}|^{(T+b)/2}\{\prod_{1\leq i<j\leq p}(\lambda_i-\lambda_j)\}^c} \operatorname{etr}\left\{-\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}(\widehat{\mathbf{\Phi}})\right\}d\mathbf{\Sigma}$$

$$= \quad C_5 \int\int\left\{\prod_{1\leq i<j\leq p}\cos^{j-i-1}(o_{ij})\right\}\frac{\lambda_p^{\frac{J-a}{2}}\left\{\prod\limits_{1\leq i<j\leq p}(\lambda_i-\lambda_j)\right\}^{1-c}}{\prod\limits_{i=1}^{p}\lambda_i^{(T+b)/2}}\operatorname{etr}\left\{-\frac{\mathbf{\Lambda}^{-1}\mathbf{OS}(\widehat{\mathbf{\Phi}})\mathbf{O}'}{2}\right\}d\mathbf{\lambda}\,d\mathbf{o}$$

$$\leq \quad C_6 \int \frac{\lambda_p^{\frac{J-a}{2}}\left\{\prod_{1\leq i<j\leq p}(\lambda_i-\lambda_j)\right\}^{1-c}}{\prod_{i=1}^{p}\lambda_i^{(T+b)/2}}\exp\left(-\sum_{j=1}^{p}\frac{\eta_p}{2\lambda_j}\right)d\mathbf{\lambda}, \tag{49}$$

where the equality follows from the transformation from $\mathbf{\Sigma}$ to $(\mathbf{\lambda}, \mathbf{o})$ as in the proof of Theorem 1.

If $c = 0$, the right hand side of (49) is bounded by

$$C_6 \int\left(\prod_{i=1}^{p}\lambda_i^{p-i}\right)\lambda_p^{(J-a)/2} = \prod_{i=1}^{p}\frac{1}{\lambda_i^{(T+b)/2}}\exp\left(-\sum_{j=1}^{p}\frac{\eta_p}{2\lambda_j}\right)d\mathbf{\lambda}$$

$$= \quad C_6\left\{\prod_{i=1}^{p-1}\int_0^\infty \frac{1}{\lambda_i^{(T+b-2p+2i)/2}}\exp\left(-\frac{\eta_p}{2\lambda_i}\right)d\lambda_i\right\}\int_0^\infty \frac{1}{\lambda_p^{(T+b-J+a)/2}}\exp\left(-\frac{\eta_p}{2\lambda_p}\right)d\lambda_p.$$

So the right hand side is integrable under Condition (B0).

If $c = 1$, the right hand side of (49) equals to

$$C_6 \int \lambda_p^{(J-a)/2}\prod_{i=1}^{p}\frac{1}{\lambda_i^{(T+b)/2}}\exp\left(-\sum_{j=1}^{p}\frac{\eta_p}{2\lambda_j}\right)d\mathbf{\lambda}$$

$$= \quad C_6\left\{\prod_{i=1}^{p-1}\int_0^\infty \frac{1}{\lambda_i^{(T+b)/2}}\exp\left(-\frac{\eta_p}{2\lambda_i}\right)d\lambda_i\right\}\int_0^\infty \frac{1}{\lambda_p^{(T+b-J+a)/2}}\exp\left(-\frac{\eta_p}{2\lambda_p}\right)d\lambda_p.$$

The right hand side is integrable under Condition (B1). The results then follow.

Appendix C: Proof for Theorem 5

Since the posterior is proper from the assumptions, it is enough to show that

$$\int\int_{\mathbb{R}^J}\|\boldsymbol{\phi}\|^2\{\operatorname{tr}(\mathbf{\Sigma}^2)\}^{h/2}L(\boldsymbol{\phi},\mathbf{\Sigma})\pi_{(0,b,c)}(\boldsymbol{\phi},\mathbf{\Sigma})d\boldsymbol{\phi}\,d\mathbf{\Sigma} < \infty. \tag{50}$$

Since $(\boldsymbol{\phi} \mid \mathbf{\Sigma}, \mathbf{Y}) \sim N_J(\hat{\boldsymbol{\phi}}, \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1})$, we have

$$\mathbb{E}(\|\boldsymbol{\phi}\|^2 \mid \mathbf{\Sigma}, \mathbf{Y}) = \mathbb{E}(\boldsymbol{\phi}'\boldsymbol{\phi} \mid \mathbf{\Sigma}, \mathbf{Y}) = \operatorname{tr}\{\mathbb{E}(\boldsymbol{\phi}\boldsymbol{\phi}' \mid \mathbf{\Sigma}, \mathbf{Y})\}$$

$$= \operatorname{tr}\{\hat{\boldsymbol{\phi}}\hat{\boldsymbol{\phi}}' + \mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}\} = \hat{\boldsymbol{\phi}}'\hat{\boldsymbol{\phi}} + \operatorname{tr}(\mathbf{\Sigma})\operatorname{tr}\{(\mathbf{X}'\mathbf{X})^{-1}\}.$$

The marginal posterior of $\mathbf{\Sigma}$ given $\mathbf{Y}$ has the form

$$m(\mathbf{\Sigma} \mid \mathbf{Y}) = C_7 \int L(\boldsymbol{\phi}, \mathbf{\Sigma})d\boldsymbol{\phi}\,\pi_{(b,c)}(\mathbf{\Sigma})$$

$$= C_8 \frac{|\mathbf{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}|^{1/2}}{|\mathbf{\Sigma}|^{(T+b)/2}\{\sum_{1\leq i<j\leq p}(\lambda_i-\lambda_j)\}^c}\operatorname{etr}\left\{-\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}(\widehat{\mathbf{\Phi}})\right\}$$

$$= C_9 \frac{1}{|\mathbf{\Sigma}|^{(T+b-Lp-1)/2}\{\sum_{1\leq i<j\leq p}(\lambda_i-\lambda_j)\}^c}\operatorname{etr}\left\{-\frac{1}{2}\mathbf{\Sigma}^{-1}\mathbf{S}(\widehat{\mathbf{\Phi}})\right\},$$

where we use the fact that $|\boldsymbol{\Sigma} \otimes (\mathbf{X}'\mathbf{X})^{-1}|^{1/2} = |\boldsymbol{\Sigma}|^{(Lp+1)/2}|\mathbf{X}'\mathbf{X}|^{-p/2}$. Therefore the left hand side of (50) equals $J_1 + J_2$, where,

$$
J_1 = C_{10} \int \frac{\{\mathrm{tr}(\boldsymbol{\Sigma}^2)\}^{h/2}}{|\boldsymbol{\Sigma}|^{(T+b-Lp-1)/2}\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^c} \,\mathrm{etr}\Big\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\Big\}d\boldsymbol{\Sigma},
$$

$$
J_2 = C_{11} \int \frac{\mathrm{tr}(\boldsymbol{\Sigma})\,\{\mathrm{tr}(\boldsymbol{\Sigma}^2)\}^{h/2}}{|\boldsymbol{\Sigma}|^{(T+b-Lp-1)/2}\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^c} \,\mathrm{etr}\Big\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\Big\}d\boldsymbol{\Sigma}.
$$

Note that $\{\mathrm{tr}(\boldsymbol{\Sigma}^2)\}^{h/2} = \{\sum_{i=1}^p \lambda_i^2\}^{h/2} \le (p\lambda_1)^h$. It easy to show that

$$
J_1 \le C_{12} \int \frac{\lambda_1^h}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^c} \,\mathrm{etr}\Big\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\Big\}d\boldsymbol{\Sigma}
$$

$$
\le C_{13} \int \frac{\lambda_1^h\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^{1-c}}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}} \exp\Big(-\sum_{i=1}^p \frac{\eta_p}{2\lambda_i}\Big)d\boldsymbol{\lambda}.
$$

If $c = 0$,

$$
J_1 \le C_{14} \int \frac{1}{\lambda_1^{(T+b-2h-2p-Lp+1)/2}} \exp\Big(-\frac{\eta_p}{2\lambda_1}\Big)d\lambda_1 \prod_{i=2}^p \int \frac{1}{\lambda_i^{(T+b-Lp-2p+2i-1)/2}} \exp\Big(-\frac{\eta_p}{2\lambda_i}\Big)d\lambda_i,
$$

which is finite if $T > (L+2)p + 2h - b + 1$. If $c = 1$,

$$
J_1 \le C_{15} \int \frac{1}{\lambda_1^{(T+b-2h-Lp-1)/2}} \exp\Big(-\frac{\eta_p}{2\lambda_1}\Big)d\lambda_1 \prod_{i=2}^p \int \frac{1}{\lambda_i^{(T+b-Lp-1)/2}} \exp\Big(-\frac{\eta_p}{2\lambda_i}\Big)d\lambda_i,
$$

which is finite if $T > Lp + 2h - b + 3$.

Similarly,

$$
J_2 \le C_{16} \int \frac{\lambda_1^{h+1}}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^c} \,\mathrm{etr}\Big\{\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\Big\}d\boldsymbol{\Sigma}
$$

$$
\le C_{17} \int \frac{\lambda_1^{h+1}\{\sum_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^{1-c}}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}} \exp\Big(-\sum_{j=1}^p \frac{\eta_p}{2\lambda_j}\Big)d\boldsymbol{\lambda}.
$$

If $c = 0$,

$$
J_2 \le C_{18} \int \frac{\lambda_1^{h+1}\prod_{i=1}^p \lambda_i^{p-i}}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}} \exp\Big(-\sum_{j=1}^p \frac{\eta_p}{2\lambda_j}\Big)d\boldsymbol{\lambda},
$$

which is finite if $T > (L+2)p + 2h - b + 3$. If $c = 1$,

$$
J_2 \le C_{19} \int \frac{\lambda_1^{h+1}}{\prod_{i=1}^p \lambda_i^{(T+b-Lp-1)/2}} \exp\Big(-\sum_{j=1}^p \frac{\eta_p}{2\lambda_j}\Big)d\boldsymbol{\lambda},
$$

which is finite if $T > Lp + 2h - b + 5$. Note that the conditions with respect to J_2 (for $c = 0, 1$) are stronger than those with respect to J_1. The theorem follows.

Appendix D: Proof for Theorem 7

The condition (AM) implies (A), (B0M) implies (B0), and (B1M) implies (B1). Thus the corresponding posteriors are all proper. It is then enough to show

$$\int \int_{\mathbb{R}^J} \|\boldsymbol{\phi}\|^k \{\text{tr}(\boldsymbol{\Sigma}^2)\}^{h/2} L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \pi_{(a,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) d\boldsymbol{\phi} \, d\boldsymbol{\Sigma} < \infty.$$

Since $\text{tr}(\boldsymbol{\Sigma}^2) = \text{tr}(\boldsymbol{\Lambda}^2) = \sum_{i=1}^p \lambda_i^2 \le p\lambda_1^2$, it is equivalent to show that

$$\int \int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \frac{1}{\|\boldsymbol{\phi}\|^{a-k} |\boldsymbol{\Sigma}|^{b/2+h} \{\prod_{1 \le i < j \le p} (\lambda_i - \lambda_j)\}^c} d\boldsymbol{\phi} \, d\boldsymbol{\Sigma}$$

$$\le \quad C_{20} \int \int_{\mathbb{R}^J} L(\boldsymbol{\phi}, \boldsymbol{\Sigma}) \lambda_1^h \pi_{(a-k,b,c)}(\boldsymbol{\phi}, \boldsymbol{\Sigma}) d\boldsymbol{\phi} \, d\boldsymbol{\Sigma} < \infty. \tag{51}$$

Since $a - k > 0$, we apply (48) to the inner integral by replacing a by $a - k$. The right hand side of (51) is bounded by

$$C_{21} \int \frac{\lambda_1^h \lambda_p^{(J-a+k)/2}}{|\boldsymbol{\Sigma}|^{(T+b)/2} \{\prod_{1 \le i < j \le p} (\lambda_i - \lambda_j)\}^c} \text{etr}\left\{-\frac{1}{2}\boldsymbol{\Sigma}^{-1}\mathbf{S}(\widehat{\boldsymbol{\Phi}})\right\} d\boldsymbol{\Sigma}$$

$$\le \quad C_{22} \int \frac{\lambda_1^h \lambda_p^{(J-a+k)/2} \{\prod_{1 \le i < j \le p}(\lambda_i - \lambda_j)\}^{1-c}}{\prod_{i=1}^p \lambda_i^{(T+b)/2}} \exp\left(-\sum_{i=1}^p \frac{\eta_p}{2\lambda_i}\right) d\boldsymbol{\lambda}. \tag{52}$$

If $c = 0$, the right hand side of (52) equals

$$C_{23} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_1})}{\lambda_1^{(T+b-2p-2h+2)/2}} d\lambda_1 \left\{\prod_{i=2}^{p-1} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_i})}{\lambda_i^{(T+b-2p+2i)/2}} d\lambda_i\right\} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_p})}{\lambda_p^{(T+b-J+a-k)/2}} d\lambda_p,$$

which is finite under Condition (B0M).

If $c = 1$, the right hand side of (52) equals

$$C_{24} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_1})}{\lambda_1^{(T+b-2h)/2}} d\lambda_1 \left\{\prod_{i=2}^{p-1} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_i})}{\lambda_i^{(T+b)/2}} d\lambda_i\right\} \int_0^\infty \frac{\exp(-\frac{\eta_p}{2\lambda_p})}{\lambda_p^{(T+b-J+a-k)/2}} d\lambda_p,$$

which is finite under Condition (B1M).

Tables on MCMC Results of Numerical Examples

Notations for Tables 1-9

$R_1(\widehat{\boldsymbol{\Sigma}})$ is the estimated frequentist risk of the Bayesian estimator for $\boldsymbol{\Sigma}$ under loss L_1 (with frequentist standard errors of the losses in parentheses).

$R_2(\widehat{\boldsymbol{\Phi}})$ is the the estimated frequentist risk of the Bayesian estimator for $\boldsymbol{\Phi}$ under loss L_2 (with frequentist standard errors of the losses in parentheses).

R_{22} is part of the R_2 associated with the non-intercept elements of $\boldsymbol{\Phi}$ (with frequentist standard errors of the losses in parentheses).

R_{Imp} is the frequentist average of sum of estimation errors of the impulse responses, as defined by (33) in the text.

Improvement in Forecast: Percentage improvement in Mean Square of One-step Forecast Errors attributable to deviation of estimates for $\boldsymbol{\Phi}$ from the true parameter relative to the MLE by Bayesian estimators. W_i, the ith element is the bracket, corresponds to percentage improvement of the ith variable by the Bayesian estimators.

Bayesian estimators based on competing priors are denoted as

CA: Bayesian estimator with constant-RATS prior;

CJ: Bayesian estimator with constant-Jeffreys prior;

CR: Bayesian estimator with constant-reference prior;

SA: Bayesian estimator with shrinkage-RATS prior;

SJ: Bayesian estimator with shrinkage-Jeffreys prior;

SR: Bayesian estimator with shrinkage-reference prior;

TMR: Bayesian estimator with the tight Minnesota-reference prior defined in the text;

LMR: Bayesian estimator with the loose Minnesota-reference prior defined in the text.

Table 1: Example 1.

	$R_1(\widehat{\boldsymbol{\Sigma}})$	$R_2(\widehat{\boldsymbol{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast (W_1, W_2)
MLE	.526(.519)	5.491(8.794)	.393(.288)	.610	
CA	.353(.382)	5.491(8.787)	.393(.288)	.611	$(-.02,\ 0.01)$
CJ	.244(.257)	5.490(8.793)	.393(.288)	.616	$(-.00, -.00)$
CR	.167(.208)	5.493(8.804)	.393(.288)	.613	$(0.00,\ 0.00)$
SA	.353(.382)	2.509(4.231)	.301(.222)	.581	$(20.76, 27.08)$
SJ	.244(.258)	2.216(3.647)	.293(.215)	.578	$(21.52, 28.54)$
SR	.161(.202)	2.005(2.879)	.287(.210)	.575	$(22.65, 29.71)$
TMR	.136(.169)	.555(.428)	.053(.027)	.456	$(78.44, 79.24)$
LMR	.157(.199)	1.199(.763)	.222(.173)	.569	$(36.65, 40.73)$

Table 2: Example 2.

	$R_1(\widehat{\boldsymbol{\Sigma}})$	$R_2(\widehat{\boldsymbol{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast (W_1, W_2)
MLE	.361(.429)	4.583(8.133)	.292(.402)	.482	
CA	.250(.312)	4.584(8.136)	.292(.402)	.494	$(0.05, 0.00)$
CJ	.202(.214)	4.584(8.124)	.292(.402)	.521	$(-.03, -.09)$
CR	.191(.185)	4.583(8.140)	.292(.402)	.521	$(0.16, -.04)$
SA	.250(.311)	1.577(1.871)	.236(.309)	.436	$(20.77, 23.90)$
SJ	.202(.213)	1.454(1.561)	.231(.296)	.476	$(22.23, 24.59)$
SR	.187(.183)	1.363(1.372)	.230(.298)	.464	$(23.55, 24.59)$
TMR	.195(.142)	.720(.641)	.096(.033)	.459	$(6.80, 53.39)$
LMR	.186(.180)	1.031(.801)	.175(.228)	.470	$(31.70, 32.85)$

Table 3: Example 3.

	$R_1(\widehat{\boldsymbol{\Sigma}})$	$R_2(\widehat{\boldsymbol{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast (W_1, W_2)
MLE	.338(.410)	2.376(3.006)	.169(.203)	.308	
CA	.236(.299)	2.376(3.009)	.169(.203)	.315	$(0.03, -.03)$
CJ	.198(.206)	2.377(3.006)	.169(.203)	.340	$(-.02, -.01)$
CR	.194(.186)	2.375(2.999)	.1687(.203)	.332	$(0.08, -.01)$
SA	.236(.300)	1.091(1.129)	.131(.154)	.281	$(18.02, 24.81)$
SJ	.198(.207)	1.052(1.002)	.125(.147)	.337	$(18.74, 25.93)$
SR	.188(.184)	1.009(.968)	.127(.150)	.300	$(19.32, 26.33)$
TMR	.583(.261)	1.860(.620)	1.311(.075)	.440	$(-205.0, -82.78)$
LMR	.188(.176)	.869(.746)	.128(.141)	.298	$(21.39, 28.27)$

Table 4: Example 4.

	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast (W_1, W_2)
MLE	.315(.331)	4.017(4.822)	.370(.407)	.099	
CA	.220(.232)	4.021(4.827)	.370(.408)	.105	$(-.06, -.07)$
CJ	.189(.157)	4.016(4.826)	.370(.406)	.114	$(-.07, -.04)$
CR	.184(.145)	4.016(4.822)	.370(.406)	.112	$(0.03, 0.14)$
SA	.220(.232)	4.896(3.495)	.371(.350)	.147	$(-26.21, -17.16)$
SJ	.189(.157)	5.426(3.496)	.376(.339)	.177	$(-33.69, -21.51)$
SR	.184(.135)	5.455(3.407)	.371(.335)	.177	$(-35.95, -19.89)$
TMR	.261(.149)	11.667(.735)	.486(.047)	.383	$(-108.0, -90.00)$
LMR	.196(.136)	9.460(1.498)	.406(.226)	.267	$(-73.75, -40.34)$

Table 5: Example 5.

	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast (W_1, W_2)
MLE	1.052(.974)	43.316(65.798)	.999(.774)	.853	
CA	.735(.759)	43.318(65.831)	.999(.774)	.873	$(-.04, .06)$
CJ	.360(.416)	43.301(65.765)	.999(.773)	.942	$(.02, -.03)$
CR	.257(.343)	43.303(65.701)	.999(.774)	.922	$(.16, .08)$
SA	.735(.760)	8.187(17.881)	.815(.624)	.733	$(20.77, 23.90)$
SJ	.360(.416)	5.352(1.055)	.816(.617)	.751	$(30.32, 27.85)$
SR	.199(.281)	3.152(3.327)	.804(.610)	.728	$(28.48, 25.76)$
TMR	.586(.407)	4.974(2.998)	1.901(.273)	.711	$(-259.9, -241.3)$
LMR	.191(.255)	2.147(1.314)	.670(.468)	.717	$(33.52, 31.75)$

Table 6: Example 6.

	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast $(W_1, W_2, W_3, W_4, W_5, W_6)$
MLE	1.410(.553)	15.681(13.483)	1.083(.306)	.353	
CA	.963(.412)	15.682(13.488)	1.082(.306)	.355	$(0.03, \ 0.01, \ -.04, \ -.01, \ -.02, \ -.09)$
CJ	.677(.282)	15.682(13.491)	1.083(.306)	.358	$(0.07, \ 0.02, \ 0.07, \ -.08, \ -.06, \ 0.10)$
CR	.255(.173)	15.684(13.483)	1.083(.305)	.354	$(0.04, \ -.08, \ -.05, \ 0.01, \ 0.03, \ 0.05)$
SA	.963(.412)	1.593(.494)	.872(.240)	.334	$(8.96, \ 14.13, \ 9.70, \ 13.67, \ 9.28, \ 7.83)$
SJ	.677(.282)	1.455(.406)	.881(.238)	.334	$(7.19, \ 12.31, \ 8.32, \ 12.45, \ 7.38, \ 5.95)$
SR	.225(.152)	1.258(.287)	.877(.240)	.329	$(7.79, \ 12.99, \ 8.25, \ 12.39, \ 8.29, \ 6.84)$
TMR	.220(.148)	2.653(.896)	.342(.068)	.312	$(64.05, 63.97, 63.68, 63.42, 64.92, 64.79)$
LMR	.250(.166)	4.369(1.404)	.830(.236)	.344	$(17.44, 21.88, 17.40, 18.54, 19.23, 18.53)$

Table 7: Example 7.

	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast $(W_1, W_2, W_3, W_4, W_5, W_6)$
MLE	.939(.392)	6.950(9.272)	1.563(.809)	.412	
CA	.666(.286)	6.953(9.269)	1.563(.809)	.423	$(0.05, -.05, -.02, -.16, -.05, -.04)$
CJ	.554(.199)	6.948(9.253)	1.563(.809)	.444	$(0.04, 0.16, 0.12, 0.16, -.08, -.02)$
CR	.459(.166)	6.952(9.298)	1.563(.809)	.421	$(-.22, 0.05, -.09, 0.03, -.06, 0.01)$
SA	.666(.286)	5.556(.596)	.657(.212)	.380	$(-17.19, 6.07, 10.51, 15.36, 2.21, 9.89)$
SJ	.554(.198)	5.650(.537)	.640(.203)	.404	$(-19.96, 4.91, 10.09, 14.30, 18.25, 5.75)$
SR	.439(.151)	5.710(.470)	.633(.200)	.396	$(-24.68, 4.49, 10.25, 14.86, 18.62, 6.47)$
TMR	.498(.140)	3.023(1.390)	.864(.146)	.454	$(-69.28, -47.19, -35.09, -20.04, -4.19, 41.58)$
LMR	.452(.165)	3.735(1.656)	1.037(.423)	.380	$(7.37, 17.40, 16.29, 19.62, 23.21, 22.15)$

Table 8: Example 8.

|	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast $(W_1, W_2, W_3, W_4, W_5, W_6)$
MLE	1.718(.692)	26.867(41.701)	3.812(1.411)	2.755	
CA	1.187(.528)	26.852(41.659)	3.8116(1.413)	2.732	$(-.07, -.07, -.14, -.06, -.11, -.06)$
CJ	.679(.258)	26.878(41.701)	3.814(1.414)	2.781	$(-.11, -.08, .15, .13, .06, -.02)$
CR	.561(.212)	26.874(41.784)	3.812(1.412)	2.709	$(0.14, 0.05, 0.19, 0.02, 0.04, 0.19)$
SA	1.187(.529)	6.607(1.117)	2.239(.607)	2.103	$(3.48, 11.83, 14.41, 2.33, 15.69, 6.61)$
SJ	.679(.259)	6.778(.958)	2.211(.575)	2.242	$(-.63, 10.15, 12.81, 17.55, 10.43, -4.02)$
SR	.528(.189)	6.795(.893)	2.209(.572)	2.167	$(-1.00, 10.48, 12.95, 17.02, 10.01, -3.43)$
TMR	1.258(.389)	5.163(1.236)	4.032(.412)	3.083	$(-1115, -754, -830, -828, -871, -746)$
LMR	.538(.199)	5.108(1.856)	2.242(.646)	2.225	$(16.48, 19.10, 19.09, 23.08, 21.76, 20.92)$

Table 9: Example 9.

|	$R_1(\widehat{\mathbf{\Sigma}})$	$R_2(\widehat{\mathbf{\Phi}})$	R_{22}	R_{Imp}	Improvement in Forecast $(W_1, W_2, W_3, W_4, W_5, W_6)$
MLE	.674(.280)	24.313(16.011)	15.662(12.752)	.143	
CA	.502(.207)	24.316(16.027)	15.668(12.772)	.150	$(0.03, -0.05, -0.09, 0.02, 0.01, 0.01)$
CJ	.436(.149)	24.309(16.012)	15.656(12.739)	.159	$(-0.13, 0.02, -0.04, 0.04, -0.06, -0.01)$
CR	.420(.149)	24.325(15.997)	15.669(12.739)	.151	$(-0.06, -0.01, 0.11, 0.11, -0.06, -0.10)$
SA	.502(.207)	13.578(1.445)	3.799(.770)	.122	$(18.02, 6.38, -50.08, 30.47, 8.92, -9.66)$
SJ	.436(.149)	14.094(1.298)	3.989(.706)	.128	$(15.45, 3.78, -57.77, 29.27, 8.87, -13.34)$
SR	.434(.142)	14.824(.986)	4.189(.609)	.130	$(11.94, 1.91, -72.42, 28.69, 10.09, -22.08)$
TMR	.488(.146)	12.658(1.564)	4.634(.418)	.128	$(47.35, 24.12, -63.63, 33.12, -8.23, -51.39)$
LMR	.420(.147)	11.542(3.608)	5.375(3.093)	.130	$(22.66, 18.67, -6.36, 29.81, 13.02, 10.48)$

(a) MLE vs. CA prior
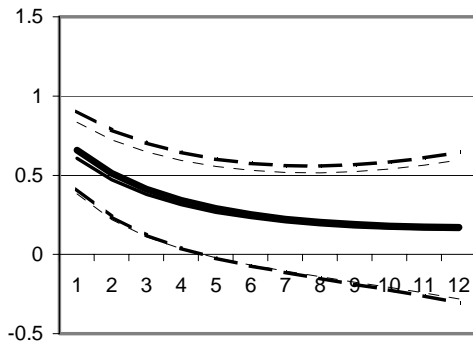
(b) MLE vs. CJ prior
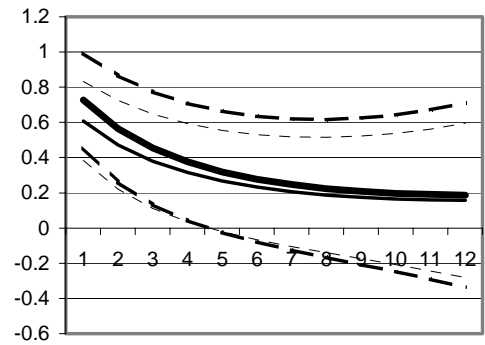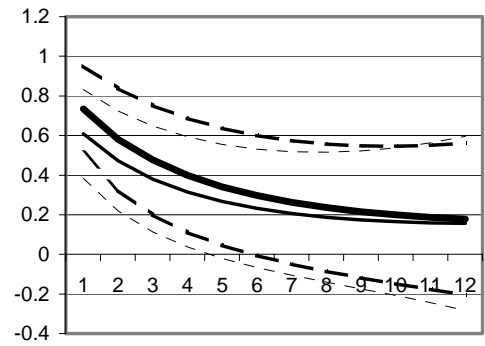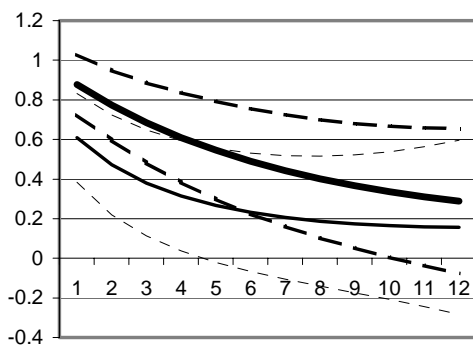
(c) MLE vs. CR prior
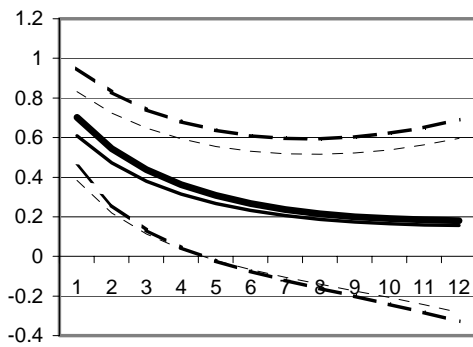
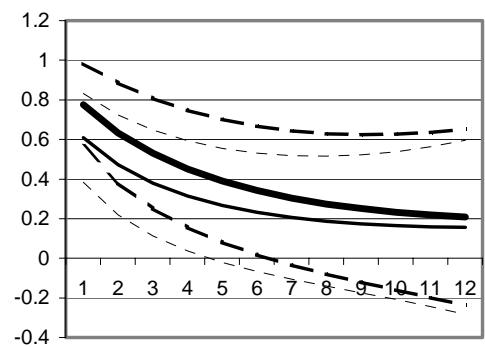(d) MLE vs. SA prior
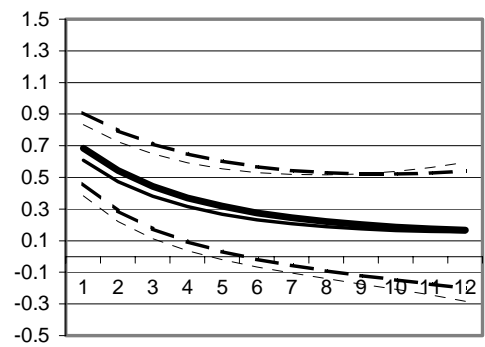
(e) MLE vs. SJ prior
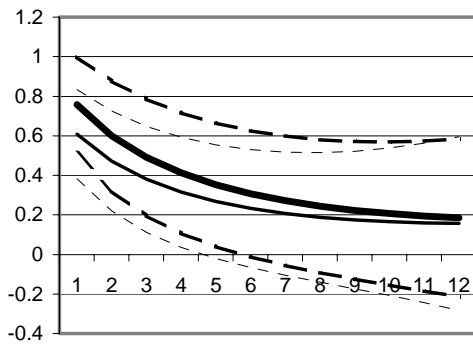
(f) MLE vs. SR prior
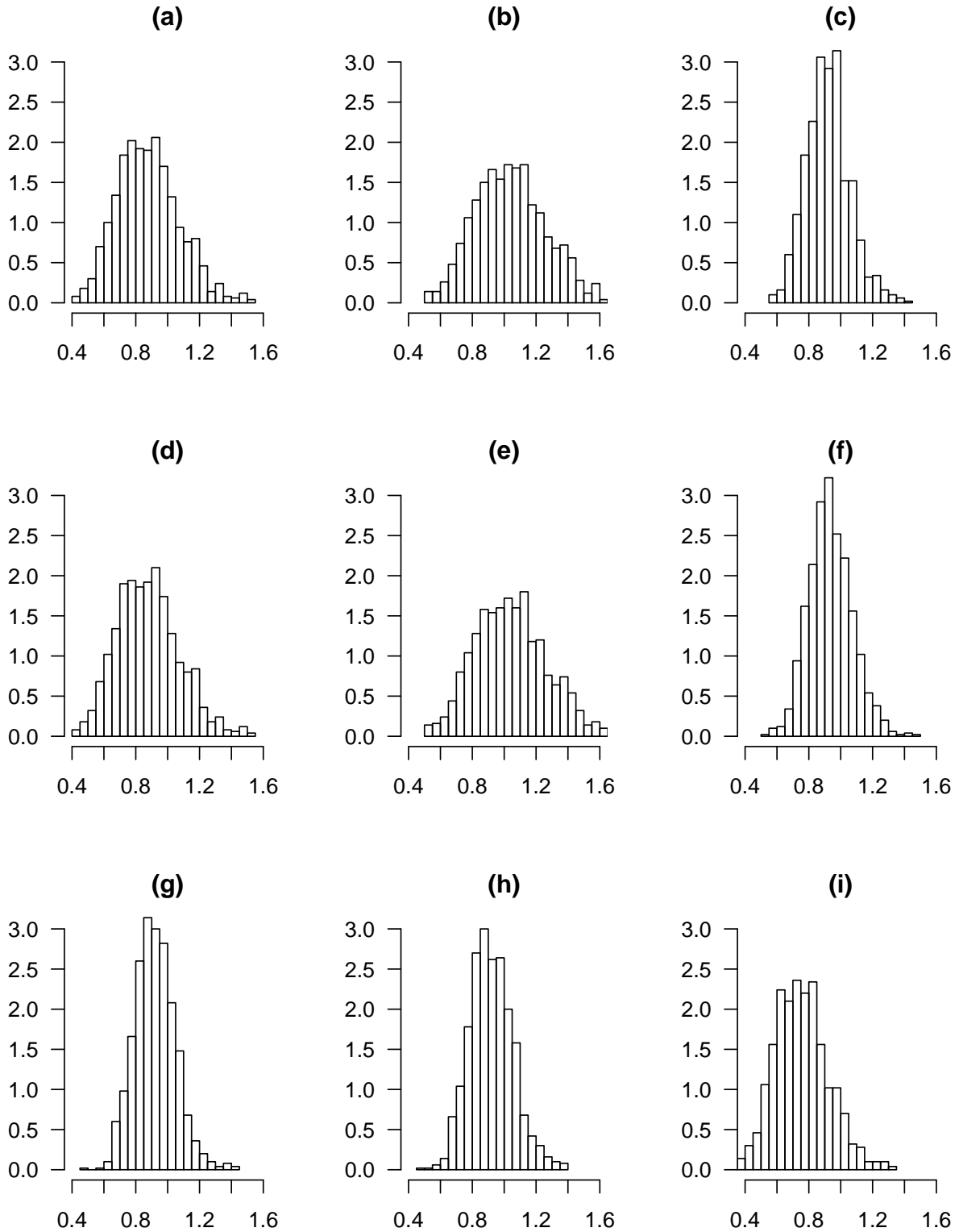
(g) MLE vs. TMR prior

(h) MLE vs. LMR prior

Figure 2: Frequentist histograms of the estimators of $\sigma_{1,1}$ in Example 6 with $p = 6, L = 1, T = 50$ and $\mathbf{\Sigma} = \mathbf{I}_6$: (a) posterior mean based on constant-RATS prior; (b) posterior mean based on constant-Jeffreys prior; (c) posterior mean based on constant-reference prior; (d) posterior mean based on shrinkage-RATS prior; (e) posterior mean based on shrinkage-Jeffreys prior; (f) posterior mean based on shrinkage-reference prior. (g) posterior mean based on a Tight Minnesota-reference prior; (h) posterior mean based on a Loose Minnesota-reference prior; (i) MLE.