Bayesian Estimator of Vector-Autoregressive Model Under the Entropy Loss

Shawn Ni, Department of Economics, University of Missouri, Columbia, MO 65211, U.S.A.

Dongchu Sun, Department of Statistics, University of Missouri, Columbia, MO 65211, U.S.A.

Abstract

The present study makes two contributions to the Bayesian Vector-Autoregression (VAR) literature. The first contribution is derivation of the Bayesian VAR estimator under the intrinsic entropy loss. The Bayesian estimator, which is distinctly different from the posterior mean, involves the frequentist expectation of a function of VAR variables. We find that the condition that allows for a closed-form expression of the frequentist expectation is violated even when the VAR is stationary, making it difficult to compute the Bayesian estimates via standard Markov Chain Monte Carlo (MCMC) procedures. The second contribution of the paper concerns MCMC simulation of the Bayesian estimator without using the closed-form expression of the frequentist expectation. A novelty of our MCMC algorithms is that they jointly simulate the posteriors of frequentist moments of VAR variables as well as the posteriors of VAR parameters. Numerical simulations show that the algorithms are surprisingly efficient.

KEY WORDS: Bayesian VAR, Entropy Loss, Latent Parameters, Markov Chain Monte Carlo.

1 Introduction

The present paper concerns Bayesian estimation of Vector-Autoregressive (VAR) models under an intrinsic entropy loss function. In the past two decades VAR has become a popular tool for modeling time series data, especially in the field of macroeconomics. A VAR of a p dimensional row variable, $\mathbf{y}_t$, $(t = 1, .., T)$ has the form $\mathbf{y}_t = \mathbf{c} + \sum_{j=1}^{L} \mathbf{y}_{t-j}\mathbf{B}_j + \boldsymbol{\epsilon}_t$, for $t = 1, \cdots, T$, where VAR lag length L is a known positive integer, $\mathbf{c}$ is a $1 \times p$ unknown vector, and $\mathbf{B}_j$ is an unknown $p \times p$ matrix. VAR residuals $\boldsymbol{\epsilon}_1, \cdots, \boldsymbol{\epsilon}_T$ are assumed to be independently identically distributed $N_p(\mathbf{0}, \boldsymbol{\Sigma})$ errors, where the covariance matrix of the error term $\boldsymbol{\Sigma}$ is an unknown $p \times p$ positive definite matrix. We now denote $\mathbf{x}_t = (1, \mathbf{y}_{t-1}, \cdots, \mathbf{y}_{t-L})$,

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

The VAR coefficient matrix $\boldsymbol{\Phi}$ is a $(1 + Lp) \times p$ matrix of unknown parameters. Then the VAR can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\Phi} + \boldsymbol{\epsilon}. \tag{1}$$

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

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

Here and hereafter $etr(\mathbf{A})$ is $\exp(trace(\mathbf{A}))$ of matrix $\mathbf{A}$. The Maximum Likelihood Estimator (MLE) of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ are [1]

$$\widehat{\boldsymbol{\Phi}}_{MLE} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}, \text{ and } \widehat{\boldsymbol{\Sigma}}_{MLE} = (\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Phi}}_{MLE})'(\mathbf{Y} - \mathbf{X}\widehat{\boldsymbol{\Phi}}_{MLE})/T \tag{3}$$

respectively.

In a typical macroeconomic application, the number of parameters to be estimated is large relative to data observations. A VAR with $p = 8$ and $L = 4$ involves 300 parameters to be estimated in $\boldsymbol{\Sigma}$ and $\boldsymbol{\Phi}$. As a result of the 'over-parameterization', the MLE $(\widehat{\boldsymbol{\Sigma}}_{MLE}, \widehat{\boldsymbol{\Phi}}_{MLE})$ for

[1] We assume that $T \geq Lp + p + 1$ and that the MLEs of $\boldsymbol{\Phi}$ and $\boldsymbol{\Sigma}$ exist.

finite samples is often found to change significantly when data sample is extended by a few more observations, when the lag L is changed slightly, or when the model incorporates a different variable. In addition, drawing finite sample inferences of the VAR parameters is a challenge. With a large number of parameters and limited data observations, asymptotic asymptotic theories may not be good guidance for finite sample properties. On the other hand, frequentist finite sample distribution is not available in analytical form for the VAR model.

In practice, Bayesian procedures are widely used for finite sample inferences of VAR models and often produce estimators with superior frequentist properties than the MLE. The Bayesian approach combines prior information with sample information to form posteriors and derive estimators from minimization of expected posterior loss. Bayesian estimators depend on researcher's choice of loss function, with tractability often being the top selection criterion. In applications of Bayesian procedures, estimators of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ are usually derived independently. For example, the posterior mean of ($\mathbf{\Sigma}$, $\mathbf{\Phi}$), seemingly the most natural Bayesian estimator, can be justified by a parametric loss function that is the sum of two separable losses–a loss with respect to $\mathbf{\Phi}$ and a loss with respect to $\mathbf{\Sigma}$. Specifically, the loss with respect to $\mathbf{\Phi}$ is the sum of squares of estimation errors of all elements in $\mathbf{\Phi}$ weighted by constants that are independent of data. The constant weighting of estimation errors gives rise to unreasonable implications. For instance, it does not take into account of the fact that data series used in the VAR may be different in scale and volatility. While Bayesian analysis based on separable loss functions has pitfalls, little research has yet been conducted to find an alternative approach.

In this paper we suggest that Bayesian estimation of VAR be based on an intrinsic loss function. The intrinsic loss we focus on is the entropy loss. The parametric form of the entropy loss function depends on the model. The entropy loss on $(\mathbf{\Sigma}, \mathbf{\Phi})$ is non-separable in $\mathbf{\Sigma}$ and $\mathbf{\Phi}$. Using the entropy loss has two appeals. First, it serves as a metric of the estimation errors in general settings. Second, for the VAR model the entropy loss offers a more plausible measure of estimation errors that economists are concerned about. The entropy loss of ($\mathbf{\Sigma}$, $\mathbf{\Phi}$) is shown to be the sum of a loss pertaining to the covariance matrix $\mathbf{\Sigma}$ and a loss pertaining to the errors of within-sample forecasts normalized by the estimated $\mathbf{\Sigma}$.

Deriving the Bayesian estimator under the entropy loss involves computing frequentist moments of VAR variables. We find that these moments of the VAR can be computed in closed form under a restrictive condition that requires the VAR to be stationary. This condition is shown to be violated by our numerical simulations even when the data-generating VAR is stationary. In other words,

for computation of Bayesian estimates we can no longer rely on standard Monte Carlo methods that require the use of the closed-form expression of the moments of VAR variables. We propose algorithms that use generated data as latent parameters in numerical simulation of posteriors and computation of Bayesian estimators under the entropy loss. The algorithms are applicable to stationary and nonstationary VARs. Data augmentation is proposed by Tanner and Wang (1987) to alter the likelihood function for easier MCMC simulation from the posteriors of parameters of interest. In recent years, data augmentation has been used for various purposes in Bayesian literature. For examples, in the study by Otrok and Whiteman (1998) the generated latent economic indicator itself is of primary interest. In the study by Elerian et al. (2001) generated data is used to estimate stochastic differential equations from discrete sample observations. Data generation in the present study is different from the existing literature in objective and in implementation. We use the generated data to compute certain frequentist moment of the VAR variables and then simulate the joint posterior distribution of the frequentist moment with VAR parameters.

In section 2 of the paper we derive the Bayesian VAR estimator under the entropy loss function. We show that the Bayesian estimator for $\mathbf{\Sigma}$ is larger than the posterior mean. The Bayesian estimator for $\mathbf{\Phi}$, which is different from the posterior mean, involves frequentist moments of VAR variables. In this section we also discuss the issue of computing the moments of VAR variables. We find unless the VAR is highly stationary, the closed-form expression of the moments cannot be used for computation of Bayesian estimators. In section 3 we provide general MCMC algorithms that simulate posteriors of parameters as well as posteriors of frequentist moments of generated data. We lay out two options for MCMC simulations. One algorithm draws VAR parameters from posteriors conditional on sample observations as well as generated data. The other one draws parameters from posteriors conditional on sample observations alone and then adds a Metropolis-Hastings step for simulations from the full conditional density. In section 4 we discuss implementation of the general algorithms for computing estimators of ($\mathbf{\Phi}$, $\mathbf{\Sigma}$) in the VAR model. In section 5 we compare the simulation results of a numerical example using the alternative algorithms. The numerical simulations demonstrate that despite a large number of latent parameters involved, the algorithms are quite efficient. In section 6 we estimate a VAR using a set of U.S. macroeconomic data. We find that the posterior risk of the Bayesian estimate is substantially lower than that of the posterior mean. In section 7 we offer concluding remarks.

2 The Entropy Loss Function and Bayesian Estimator

A common frequentist argument against Bayesian analysis is that Bayesian estimators depend on researchers' choices of priors and loss functions and that these choices are often made without sufficient degree of generalization. Given the fact that researchers use VAR models for a variety of applications it is useful to establish a framework that allows for easier interpretations of research findings by others. An extensive literature has been developed in Bayesian statistics for selecting priors that are in some sense 'noninformative' or 'objective' to serve as a reference for inference. These objective priors are derived from certain general principles.[2] Scientific reporting is made more convenient by the common use of objective priors since they are to a large extent independent of researchers' individual preferences.

A similar argument can be made on the choice of loss function, which is a metric for the difference between the true parameter $(\mathbf{\Phi}, \mathbf{\Sigma})$ and an estimator $(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})$. One may consider a loss function that is made of separable part for $\mathbf{\Sigma}$ and $\mathbf{\Phi}$, each takes a given parametric form (e.g., a quadratic function). The overall loss with respect to $(\mathbf{\Phi}, \mathbf{\Sigma})$ is then in the form of

$$L(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) = L_1(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}) + L_2(\widehat{\mathbf{\Phi}}; \mathbf{\Phi}). \tag{4}$$

In this setting a Bayesian estimator $\widehat{\mathbf{\Sigma}}$ is selected independent of the estimator $\widehat{\mathbf{\Phi}}$. In economic applications this restriction gives rise to unreasonable results, as the ensuing discussion will illustrate. In addition, the parametric functions chosen by different researchers are ad hoc and it is difficult to argue that one choice is more reasonable than another. If in applications different parametric loss functions lead to substantially different estimates, there is no obvious criterion to select among them. Instead of relying on parametric loss functions that are based on researcher's preference, an alternative approach is to adopt an intrinsic loss function that defines a metric in a general sense, with its parametric form depending on the problem at hand. One natural choice of such metric is the entropy function.

[2]For example, the Jeffreys prior, which is proportional to the square root of the determinant of the Fisher information matrix, is derived from the "invariance principle"– meaning the prior is invariant to re-parameterization (see Jeffreys 1961 and Zellner 1971). Another class of priors are derived by maximizing the difference between information content in the posterior and prior. Zellner's Maximal Data Information prior (1971) and Bernardo's (1979) and Berger and Bernardo's (1992) reference prior are based on this approach. For a recent review of various approaches for deriving noninformative priors see Kass and Wasserman (1996). Ni and Sun (2001) compare the properties of Bayesian VAR estimators under various non-informative priors.

2.1 The entropy loss function

The general form of the entropy loss is defined in Robert (1994, p74). For the VAR model it is given by

$$\begin{aligned} L_E(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) &= \int \log \left\{ \frac{f(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}{f(\mathbf{Y}|\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})} \right\} f(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma}) d\mathbf{Y} \\ &= \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})} \log \left\{ \frac{f(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}{f(\mathbf{Y}|\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})} \right\} \end{aligned} \tag{5}$$

where f is the density of VAR variables $\mathbf{Y}$. In information theory $\log(1/f(\mathbf{Y}))$ is often used to measure the content of information regarding the VAR parameters when a researcher observes $\mathbf{Y}$. Thus the entropy loss can be interpreted as the expected difference in information gained from data observation when researcher's estimates of the VAR parameters are $(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})$ instead of the true parameters $(\mathbf{\Phi}, \mathbf{\Sigma})$. Note that for computing the frequentist expectation in the loss function, $(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})$ are not treated as functions of $\mathbf{Y}$. Naturally, the larger the entropy loss the larger the difference between $(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})$ and the true parameters $(\mathbf{\Phi}, \mathbf{\Sigma})$.

In the following, for the multivariate normal model we decompose the loss L_E into two parts, a part measures the loss associated with the covariance matrix $\mathbf{\Sigma}$ only, the second part measures the loss of VAR coefficients but is related to the covariance matrix as well as frequentist expectation $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$. Throughout the paper the notation $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}$ represents frequentist expectation given parameter $(\mathbf{\Phi}, \mathbf{\Sigma})$. In our VAR notations of the previous section $\mathbf{X}$ are the lags of $\mathbf{Y}$. We will use both $\mathbf{Y}$ and $\mathbf{X}$ as symbols of VAR variables. Note that for a finite sample $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ exists even when the VAR has explosive roots.

Lemma 1 *Denote the* $(1 + Lp) \times (1 + Lp)$ *frequentist expectation matrix as*

$$\boldsymbol{G} = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\boldsymbol{X}'\boldsymbol{X}). \tag{6}$$

The entropy loss function L_E *can be decomposed into two parts,*

$$L_E(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) = L_{E1}(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}) + L_{E2}(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}). \tag{7}$$

where

$$L_{E1}(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}) = \frac{T}{2}\{tr(\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - \log|\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}| - p\}, \tag{8}$$

$$L_{E2}(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) = \frac{1}{2} tr[\widehat{\mathbf{\Sigma}}^{-1}\{(\mathbf{\Phi} - \widehat{\mathbf{\Phi}})'\boldsymbol{G}(\mathbf{\Phi} - \widehat{\mathbf{\Phi}})\}. \tag{9}$$

Proof. Since $\mathbf{y}_t - \mathbf{X}_t\mathbf{\Phi}$, $t = 1, \cdots, T$ are iid. $\sim N_p(\mathbf{0}, \mathbf{\Sigma})$, we have $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\{\mathbf{Y} - \mathbf{X\Phi}\} = \mathbf{0}$, $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\{(\mathbf{Y} - \mathbf{X\Phi})'\mathbf{X\Phi}\} = \mathbf{0}$, and $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\{(\mathbf{Y} - \mathbf{X\Phi})'(\mathbf{Y} - \mathbf{X\Phi})\} = T\mathbf{\Sigma}$.

$$
\begin{aligned}
& \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\left\{\log\frac{|\mathbf{\Sigma}|^{-T/2}etr\{-\frac{1}{2}(\mathbf{Y}-\mathbf{X\Phi})\mathbf{\Sigma}^{-1}(\mathbf{Y}-\mathbf{X\Phi})'\}}{|\widehat{\mathbf{\Sigma}}|^{-T/2}etr\{-\frac{1}{2}(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{\Phi}})\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{\Phi}})'\}}\right\} \\
= & \frac{T}{2}(\log|\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}| - p) + \frac{1}{2}tr\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\left\{(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{\Phi}})\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{Y}-\mathbf{X}\widehat{\mathbf{\Phi}})'\right\} \\
= & \frac{T}{2}(\log|\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}| - p) + \frac{1}{2}tr(\widehat{\mathbf{\Sigma}}\mathbf{\Sigma}^{-1}T) + \frac{1}{2}tr\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}\left\{\mathbf{X}(\mathbf{\Phi}-\widehat{\mathbf{\Phi}})\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{\Phi}-\widehat{\mathbf{\Phi}})'\mathbf{X}'\right\} \\
= & \frac{T}{2}\{tr(\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - \log|\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}| - p\} + \frac{1}{2}tr\left\{(\mathbf{\Phi}-\widehat{\mathbf{\Phi}})\widehat{\mathbf{\Sigma}}^{-1}(\mathbf{\Phi}-\widehat{\mathbf{\Phi}})'\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})\right\}.
\end{aligned}
$$

The result follows. □

The next theorem presents the Bayesian estimator under the loss L_E.

Theorem 1 *Under the loss L_E, the generalized Bayesian estimator of $(\mathbf{\Phi}, \mathbf{\Sigma})$ is*

$$
\widehat{\mathbf{\Phi}}_E = \left[\mathbb{E}\{\boldsymbol{G} \mid \boldsymbol{Y}\}\right]^{-1}\mathbb{E}\{\boldsymbol{G\Phi} \mid \boldsymbol{Y}\}, \tag{10}
$$

$$
\widehat{\mathbf{\Sigma}}_E = \mathbb{E}(\mathbf{\Sigma} \mid \boldsymbol{Y}) + \frac{1}{T}\mathbb{E}\{(\mathbf{\Phi}-\widehat{\mathbf{\Phi}}_E)'\boldsymbol{G}(\mathbf{\Phi}-\widehat{\mathbf{\Phi}}_E) \mid \boldsymbol{Y}\}. \tag{11}
$$

where $\boldsymbol{G}$ is given by (6).

Proof. Let $(\widetilde{\mathbf{\Phi}}, \widetilde{\mathbf{\Sigma}})$ denote an arbitrary estimator of $(\mathbf{\Phi}, \mathbf{\Sigma})$. For the loss function L_E and posterior $\pi(\mathbf{\Phi}, \mathbf{\Sigma}|\mathbf{Y})$, the expected posterior loss is

$$
\begin{aligned}
R(\widetilde{\mathbf{\Phi}}, \widetilde{\mathbf{\Sigma}}|\mathbf{Y}) &= \mathbb{E}\Big\{L_E(\widetilde{\mathbf{\Phi}}, \widetilde{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) \mid \mathbf{Y}\Big\} \\
&= \mathbb{E}\Big\{L_{E1}(\widetilde{\mathbf{\Sigma}}; \mathbf{\Sigma}) \mid \mathbf{Y}\Big\} + \mathbb{E}\Big\{L_{E2}(\widetilde{\mathbf{\Phi}}, \widetilde{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) \mid \mathbf{Y}\Big\}.
\end{aligned}
$$

The Bayesian estimator, which minimizes the expected posterior loss, can be derived through the first order conditions. Note that for any matrices $\mathbf{A}, \mathbf{B}$, and $\mathbf{C}$,

$$
\frac{\partial tr(\mathbf{ABA'C})}{\partial \mathbf{A}} = \mathbf{CAB} + \mathbf{C'AB'}.
$$

It follows that

$$
\frac{\partial R(\widetilde{\mathbf{\Phi}}, \widetilde{\mathbf{\Sigma}}|\mathbf{Y})}{\partial \widetilde{\mathbf{\Phi}}} = \mathbb{E}\{\mathbf{G}(\mathbf{\Phi}-\widetilde{\mathbf{\Phi}}) \mid \mathbf{Y}\}\widetilde{\mathbf{\Sigma}}^{-1}. \tag{12}
$$

Setting the right hand side of Equation (12) to 0 gives the Bayesian estimator for $\mathbf{\Phi}$ in (10). Using the result that for any matrices $\mathbf{A}$ and $\mathbf{B}$,

$$\frac{\partial \log(|\mathbf{A}|)}{\partial \mathbf{A}} = (\mathbf{A}^{-1})', \ \frac{\partial tr(\mathbf{A}^{-1}\mathbf{B})}{\partial \mathbf{A}} = -(\mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{-1})',$$

we have

$$\frac{\partial}{\partial \widetilde{\mathbf{\Sigma}}} R(\widehat{\mathbf{\Phi}}_E, \widetilde{\mathbf{\Sigma}}|\mathbf{Y}) = \mathbb{E}\left\{ \frac{T}{2}(-\widetilde{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}\widetilde{\mathbf{\Sigma}}^{-1} + \widetilde{\mathbf{\Sigma}}^{-1}) - \frac{1}{2}\widetilde{\mathbf{\Sigma}}^{-1}(\mathbf{\Phi} - \widehat{\mathbf{\Phi}}_E)'\mathbf{G}(\mathbf{\Phi} - \widehat{\mathbf{\Phi}}_E)\widetilde{\mathbf{\Sigma}}^{-1} \,\middle|\, \mathbf{Y} \right\}. \quad (13)$$

Let the right hand side of Equation (13) be 0, we get

$$\mathbb{E}\left\{ \mathbf{\Sigma} - \widetilde{\mathbf{\Sigma}} + \frac{1}{T}(\mathbf{\Phi} - \widehat{\mathbf{\Phi}}_E)'\mathbf{G}(\mathbf{\Phi} - \widehat{\mathbf{\Phi}}_E) \,\middle|\, \mathbf{Y} \right\} = 0.$$

From this equation the estimator $\widehat{\mathbf{\Sigma}}_E$ can be derived. □

2.2 Comparing the Bayesian estimator with the posterior mean

Lemma 1 shows that the entropy loss is related to a separable loss function. The first part of the entropy loss L_E is a loss concerning $\mathbf{\Sigma}$ only, the second part of the entropy loss L_E is similar to a quadratic loss. Specifically, we consider the following loss functions closely related with the entropy loss: for $\mathbf{\Sigma}$ we consider a pseudo entropy loss function

$$L_1(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}) = \frac{T}{2} tr(\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - \log|\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}| - p, \quad (14)$$

where p is the number of variables in the VAR; and for $\mathbf{\Phi}$ we consider a quadratic function

$$L_2(\widehat{\mathbf{\Phi}}; \mathbf{\Phi}) = \frac{1}{2} tr\{(\widehat{\mathbf{\Phi}} - \mathbf{\Phi})'\mathbf{W}(\widehat{\mathbf{\Phi}} - \mathbf{\Phi})\}, \quad (15)$$

where $\mathbf{W}$ is a constant weighting matrix. Bayesian estimators of $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ can be derived independently from minimizing expected loss functions regarding $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ respectively. The separable loss function is associated with the posterior mean estimator. The following fact is straightforward.

Fact 1 *(a) Under the loss L_1, the generalized Bayesian estimator of $\mathbf{\Sigma}$ is $\widehat{\mathbf{\Sigma}}_{Mean} = \mathbb{E}(\mathbf{\Sigma} \mid \boldsymbol{Y})$. (b) Under the loss L_2, the generalized Bayesian estimator of $\mathbf{\Phi}$ is $\widehat{\mathbf{\Phi}}_{Mean} = \mathbb{E}(\mathbf{\Phi} \mid \boldsymbol{Y})$.*

The above fact shows that using posterior mean as the Bayesian estimator is equivalent to treating the weighting matrix $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ as constant and ignores the role played by $\widehat{\mathbf{\Sigma}}$ in L_{E2}. The loss L_2 is weighted estimation errors of all elements of $\mathbf{\Phi}$: $\frac{1}{2}\sum_{j=1}^{p}\sum_{k=1}^{1+Lp}\sum_{i=1}^{1+Lp}(\widehat{\Phi}_{k,j} - \Phi_{k,j})w_{k,i}(\widehat{\Phi}_{i,j} - \Phi_{i,j})$. If the weighting matrix $\mathbf{W}$ is the identity matrix, then the loss of L_2 is

simply the sum of squared errors of all elements of $\mathbf{\Phi}$, $\frac{1}{2}\sum_{i=1}^{1+Lp}\sum_{j=1}^{p}(\widehat{\Phi}_{i,j}-\Phi_{i,j})^2$. 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.

In contrast to the ad hoc separable loss function, the entropy loss involves a more complicated weighting scheme which is far more reasonable. Denote $\widehat{\epsilon}=\mathbf{X}\mathbf{\Phi}-\mathbf{X}\widehat{\mathbf{\Phi}}$ as the difference between estimated residuals and the true residuals, which can be interpreted as the within-sample forecast errors attributed to the estimation error of $\mathbf{\Phi}$. With this notation, L_{E2} can be rewritten as $\frac{1}{2}tr\left\{\widehat{\mathbf{\Sigma}}^{-1}\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\widehat{\epsilon}'\widehat{\epsilon})\right\}$. In loss L_{E2}, elements of estimation errors in $\mathbf{\Phi}$ are weighted according to their contributions to the covariance of forecast errors normalized by the inverse of the estimated covariance of the residuals. Therefore the entropy loss is a more natural metric for the fit of estimator in frequentist terms than the quadratic loss. Under the entropy loss the Bayesian estimator of $\mathbf{\Phi}$ is different from the posterior mean. Although L_{E1} is the same as L_1, the Bayesian estimator of $\mathbf{\Sigma}$ is different from the posterior mean because of the presence of $\widehat{\mathbf{\Sigma}}$ in L_{E2}.

Theorem 1 shows that the Bayesian estimator $\widehat{\mathbf{\Sigma}}_E$ under the intrinsic loss is strictly larger than the posterior mean. This result can be explained by the form of the entropy loss. The Bayesian estimator $\widehat{\mathbf{\Sigma}}_E$ minimizes the posterior risk by striking an optimal balance between the two parts of the loss, L_{E1} and L_{E2}. The posterior mean $\mathbb{E}(\mathbf{\Sigma}|Y)$ minimizes L_{E1}-related posterior risk with no regard to L_{E2}-related risk. The L_{E1}-related posterior risk of Bayesian estimator $\widehat{\mathbf{\Sigma}}_E$ derived in Theorem 1 is larger than that of the posterior mean. But the larger L_{E1}-related risk of the Bayesian estimator $\widehat{\mathbf{\Sigma}}_E$ is more than compensated by a smaller L_{E2}-related risk. If $\mathbf{G}$ is very large then the gain by using the Bayesian estimator in place of the posterior mean can be sizable.

To compare the Bayesian estimator $\widehat{\mathbf{\Phi}}_E$ with the posterior mean, note that

$$\begin{aligned}\widehat{\mathbf{\Phi}}_E &= \Big[\mathbb{E}\{\mathbf{G}\,|\,\mathbf{Y}\}\Big]^{-1}\mathbb{E}\{\mathbf{G}\mathbf{\Phi}\,|\,\mathbf{Y}\}\\ &= \mathbb{E}(\mathbf{\Phi}|\mathbf{Y})+\Big[\mathbb{E}\{\mathbf{G}\,|\,\mathbf{Y}\}\Big]^{-1}COV[\mathbf{G},\mathbf{\Phi}\,|\,\mathbf{Y}].\end{aligned}$$

It is likely that $\mathbf{G}=\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ and $\mathbf{\Phi}$ are positively correlated when VAR variables exhibit positive serial correlation. In this case the Bayesian estimator of $\mathbf{\Phi}$ under the entropy loss is larger than the posterior mean. It is known that the MLE of $\mathbf{\Phi}$ has a downward bias when the true parameters are closed to random walk, a typical pattern for macroeconomic data. With a flat prior on $\mathbf{\Phi}$, the posterior mean of $\mathbf{\Phi}$ is likely to show a downward bias as well. We conjecture that the

Bayesian estimator of $\mathbf{\Phi}$ based on the entropy loss may be helpful in correcting the bias in the posterior mean. Our numerical simulation results show that it is indeed the case.

2.3 Numerical simulations of $(\mathbf{\Sigma}, \mathbf{\Phi})$ using closed-form expression $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$

Suppose the posterior of $(\mathbf{\Sigma}, \mathbf{\Phi})$ is available (either as a standard distribution or as a simulated distribution), and suppose in addition the frequentist expectation $\mathbf{G} = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ is available in closed form for any given value of $(\mathbf{\Sigma}, \mathbf{\Phi})$, then the Bayesian estimator $(\mathbf{\Sigma}, \mathbf{\Phi})$ can be calculated using the result of Theorem 1.

The posterior of $(\mathbf{\Sigma}, \mathbf{\Phi})$ depends on the prior. The most popular noninformative prior for $\mathbf{\Sigma}$ is the Jeffreys prior (See Geisser 1965, Tiao and Zellner 1964). Specifically for the VAR covariance matrix, the Jeffreys prior is

$$\pi_J(\mathbf{\Sigma}) \propto \frac{1}{|\mathbf{\Sigma}|^{(p+1)/2}}. \tag{16}$$

The prior for $(\mathbf{\Phi}, \mathbf{\Sigma})$ can be obtained by putting together priors for $\mathbf{\Phi}$ and $\mathbf{\Sigma}$. In practice, it is often more convenient to consider the vectorized form $\boldsymbol{\phi} = vec(\mathbf{\Phi})$, instead of $\mathbf{\Phi}$. A common expression of ignorance about $\boldsymbol{\phi}$ is a (flat) constant prior. A popular noninformative prior for multivariate regression models consists of a constant prior for $\boldsymbol{\phi}$ and the Jeffreys prior for $\mathbf{\Sigma}$. A similar prior is used by the RATS package. The joint densities of $(\boldsymbol{\phi}, \mathbf{\Sigma})$ under the constant-Jeffreys prior are in the form

$$\pi_{CJ}(\boldsymbol{\phi}, \mathbf{\Sigma}) \propto \frac{1}{|\mathbf{\Sigma}|^{(p+1)/2}}, \tag{17}$$

and for the constant-RATS prior

$$\pi_{CA}(\boldsymbol{\phi}, \mathbf{\Sigma}) \propto \frac{1}{|\mathbf{\Sigma}|^{(L+1)p/2+1}}. \tag{18}$$

For an argument of using constant-RATS instead of constant-Jeffreys prior see Sims and Zha (1999).

In most applications of VAR models, posteriors based on commonly employed priors and data distributions are not standard distributions. In these situations the posterior distributions can be simulated using MCMC method. Besides the papers cited in the introduction, applications of Monte Carlo methods are shown to be fruitful for a variety of topics of Bayesian econometrics in the studies of Geweke (1989, 1993), Chib and Greenberg (1996), Sims and Zha (1999), and DeJong et al. (2000), among others. To illustrate a basic procedure of MCMC, we draw the VAR coefficients under the constant-RATS prior for a data sample denoted as $(\mathbf{X}, \mathbf{Y})$. Suppose

after cycle $k-1$ we have sampled $(\mathbf{\Sigma}_{k-1}, \mathbf{\Phi_{k-1}})$. In cycle k we simulate from an Inverse Wishart distribution [3] $\mathbf{\Sigma}_k \sim IW(\{\mathbf{S}(\widehat{\mathbf{\Phi}}_{MLE})\}, T)$. Then simulate $\boldsymbol{\phi}_k$ from a multivariate normal distribution $MVN(\widehat{\boldsymbol{\phi}}_{MLE}, \mathbf{\Sigma}_k \otimes (\mathbf{X}'\mathbf{X})^{-1})$. Equipped with the closed-form expression for $\mathbf{G}$, we then calculate the matrix $\mathbf{G}_k = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}_k, \mathbf{\Sigma}_k)}(\mathbf{X}'\mathbf{X})$. The Bayesian estimate of $(\mathbf{\Phi}, \mathbf{\Sigma})$ is then calculated based on the moments of posterior density using the formula given in Theorem 1. It is obvious that the Monte Carlo algorithm is applicable only when $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ can be computed throughout all MCMC cycles.

2.4 Computing the frequentist expectation $\mathbb{E}(\mathbf{X}'\mathbf{X} \,|\, \mathbf{\Phi}, \mathbf{\Sigma})$

For computation of the frequentist expectation $\mathbf{G} = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ in a stationary VAR we rewrite the VAR following the notations of Lutkepohl (1993). We define

$$\begin{aligned} \boldsymbol{\mu} &= (\mathbf{I}_p - \mathbf{B}_1 - \cdots - \mathbf{B}_L)^{-1}\mathbf{c}, \\ \tilde{\mathbf{y}}_t &= \mathbf{y}_t - \boldsymbol{\mu}, \\ \widetilde{\mathbf{Y}}_t &= (\mathbf{y}_t - \boldsymbol{\mu}, \mathbf{y}_{t-1} - \boldsymbol{\mu}, \cdots, \mathbf{y}_{t-L} - \boldsymbol{\mu}). \end{aligned}$$

Clearly $\widetilde{\mathbf{Y}}_t$ has a stationary distribution with auto-covariance matrix

$$\mathbf{\Gamma}(h) = \mathbb{E}(\widetilde{\mathbf{Y}}_t'\widetilde{\mathbf{Y}}_{t-h}).$$

We define $\mathbf{\Sigma}_u = diag(\mathbf{\Sigma}, \mathbf{0}, \cdots, \mathbf{0})_{Lp \times Lp}$ and

$$\mathbf{B} = \begin{pmatrix} \mathbf{B}_1 & \mathbf{I}_p & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{B}_2 & \mathbf{0} & \mathbf{I}_p & \cdots & \mathbf{0} \\ \cdot & \cdot & \cdot & \cdots & \cdot \\ \mathbf{B}_L & \mathbf{0} & \mathbf{0} & \cdots & \mathbf{I}_p \end{pmatrix}_{Lp \times Lp}.$$

A closed-form expression of $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi}, \mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ is available under the following condition.

Condition A: The matrix $(\mathbf{I}_{L^2p^2 \times L^2p^2} - \mathbf{B} \otimes \mathbf{B})$ is invertable and positive semi-definite, and the initial observations $\mathbf{y}_{-1}$,..., $\mathbf{y}_{-L}$ equal to the unconditional mean $\boldsymbol{\mu}$.

Lemma 2 *Assume Condition A holds then*

a) The covariance matrix $\mathbf{\Gamma}(0)$ *has the expression*

$$vec\,\{\mathbf{\Gamma}(0)\} = (\boldsymbol{I}_{L^2p^2 \times L^2p^2} - \boldsymbol{B} \otimes \boldsymbol{B})^{-1} vec\,(\mathbf{\Sigma}_u). \tag{19}$$

[3] There are more than one definition of IW distribution (e.g., see Press 1982 and Anderson 1984). We use the definition given by Anderson (1984).

(b) The matrix $\boldsymbol{G} = \mathbb{E}_{(\mathbf{Y}|\boldsymbol{\Phi},\boldsymbol{\Sigma})}(\boldsymbol{X}'\boldsymbol{X})$ *can be computed as*

$$\mathbb{E}_{(\mathbf{Y}|\boldsymbol{\Phi},\boldsymbol{\Sigma})}(\boldsymbol{X}'\boldsymbol{X}) \quad = \quad T(\mathbf{1}, \boldsymbol{\mu} \otimes \boldsymbol{I})'(\mathbf{1}, \boldsymbol{\mu} \otimes \boldsymbol{I}) + (T-1)diag(\mathbf{0}, \boldsymbol{\Gamma}(0)). \tag{20}$$

The Lemma can be proved using derivation similar to that of Lutkepohl (1993). If the second part of Condition A on initial observations is relaxed the matrix $\mathbf{G}$ can still be computed, but via a modified formula.

Condition A requires eigenvalues of the matrix $(\mathbf{I}_{L^2p^2 \times L^2p^2} - \mathbf{B} \otimes \mathbf{B})$ to be positive, which needs all roots of matrix $\mathbf{B} \otimes \mathbf{B}$ to be stationary. The condition is obviously violated if the VAR contains explosive roots, a common case for macroeconomic time series. Furthermore, to make use of the results with the closed-form expression of $\mathbb{E}_{(\mathbf{Y}|\boldsymbol{\Phi},\boldsymbol{\Sigma})}(\mathbf{X}'\mathbf{X})$ for the purpose of computing Bayesian estimators using posteriors simulated via Monte Carlo method, the simulated parameters must satisfy condition A in each MCMC cycle. Some generated VAR coefficients $\boldsymbol{\Phi}_k$ (in the kth MCMC cycle) may have explosive roots even when the true VAR parameters are stationary.

To study how restrictive the condition is for using closed-form expression of $\mathbf{G}$, consider the following example, a VAR with the true parameters

$$\boldsymbol{\Sigma} \quad = \quad \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 2.0 & 0 \\ 0 & 0 & 0 & 0 & 2.5 \end{pmatrix}, \; \boldsymbol{\Phi} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ \mathrm{b} & 0 & 0 & 0 & 0 \\ 0 & \mathrm{b} & 0 & 0 & 0 \\ 0 & 0 & \mathrm{b} & 0 & 0 \\ 0 & 0 & 0 & \mathrm{b} & 0 \\ 0 & 0 & 0 & 0 & \mathrm{b} \end{pmatrix}.$$

We conduct the following experiment. First, we generate 1000 time series with the diagonal elements of matrix $\boldsymbol{\Phi}$, b, set at a given value (to be specified later), with $\boldsymbol{\Sigma}$ given above and the sample size T set as 100. Then for each data sample we simulate 10000 MCMC cycles that generate numerical posterior distributions of ($\boldsymbol{\Phi}$,$\boldsymbol{\Sigma}$) following the procedure described in Section 2.3. For a given data set the MCMC cycles stop when some diagonal elements of covariance matrix $\boldsymbol{\Gamma}(0)$ become negative. For that data set the MCMC procedure using the closed-form expression of $\mathbf{G}$ matrix fails. We report the frequency of failure of the procedure out of 1000 data sets. The failure rate varies with the parameter b.

First, let b be 0.2. We do not find any case of negative variance in $\boldsymbol{\Gamma}(0)$. Hence with the eigenvalues of the matrix $\boldsymbol{\Phi}$ sufficiently close to zero we can use the closed-form solution to calculate the frequentist expectation $\mathbb{E}_{(\mathbf{Y}|\boldsymbol{\Phi},\boldsymbol{\Sigma})}(\mathbf{X}'\mathbf{X})$ and the Bayesian estimator under the entropy loss.

Table 1: Failure Frequency of MCMC Algorithm using the Closed-Form Expression of $\mathbf{G}$

b	Failure Frequency
0.2	0.000
0.5	0.412
0.6	0.896
0.8	1.000

We now let b be 0.5, which is far from unity. Despite the stationarity of the true VAR, it turns out that the simulated matrix $\mathbf{\Gamma}(0)$ often has at least one negative variance component. Out of 1000 generated data samples, the MCMC procedure breaks down 412 times.

We then increase the parameter b to 0.6 and repeat the experiment. Under this set of parameters, the failure incidents of the MCMC procedure using the closed-from expression of $\mathbf{G}$ are increased to 896 out of 1000 generated data sets. When b is raised to 0.8, the VAR is still stationary, but in every data sample some variance components of the simulated covariance matrix $\mathbf{\Gamma}(0)$ become negative, rendering Lemma 2 inapplicable. One may suspect that the high failure rate of the MCMC routine is caused by the large variance components in the $\mathbf{\Sigma}$ matrix of this example. It turns out not to be the case. Note that in the kth cycle of the MCMC algorithm the variance of the simulated $\boldsymbol{\phi}_k$ vector is $\mathbf{\Sigma}_k \otimes (\mathbf{X}'\mathbf{X})^{-1}$. If the variance components of the true $\mathbf{\Sigma}$ used to generate data are smaller then $(\mathbf{X}'\mathbf{X})$ is smaller and $(\mathbf{X}'\mathbf{X})^{-1}$ larger for the generated data. As a result the MCMC routine fails in approximately same frequency as reported in the above table when the variance components of $\mathbf{\Sigma}$ are substantially reduced.

The experiment shows that computation of the Bayesian estimator is possible for very limited cases if we use the closed-form expression to compute $\mathbf{G}$ matrix. Furthermore, it is not feasible to add a step to simulate the $\mathbf{G}$ matrix in the standard MCMC procedure. It is because that in the kth MCMC cycle $\mathbf{G}_k$ is the frequentist mean which is a function of parameters $(\mathbf{\Phi}_k, \mathbf{\Sigma}_k)$. If for the purpose of computing the frequentist mean $\mathbf{G}_k$ we generate 5000 data sample from the VAR with parameters $(\mathbf{\Phi}_k, \mathbf{\Sigma}_k)$ then the amount of computation required for the whole MCMC procedure is increased by the additional time needed for simulating the $\mathbf{G}_k$ matrix multiplied by the number of MCMC cycles. Suppose it takes 10 seconds to generate the 5000 VAR samples and compute the $\mathbf{G}$ matrix for each MCMC cycle and the number of MCMC cycles is 10000, then the additional amount

of time added to the standard MCMC procedure is about 28 hours for computation of Bayesian estimates for one data sample. For practical purposes, we must take an alternative approach to compute Bayesian estimates under the entropy loss.

We propose some general algorithms to deal with the difficulty raised in computing the frequentist expectation $\mathbf{G}$. Note that to compute the Bayesian estimators of $\mathbf{\Phi}$ and $\mathbf{\Sigma}$, we need compute the quantities $\mathbb{E}\{\mathbf{G} \mid \mathbf{Y}\}$ and $\mathbb{E}\{\mathbf{G\Phi} \mid \mathbf{Y}\}$. The challenge is to compute the posterior moments without first deriving the closed-form frequentist expectation $\mathbf{G} = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ as the conventional procedure requires. Our approach is to generate data as latent parameters which are used to obtain $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$. To be more specific, we define a random matrix $\mathbf{Y}^*$ which is independent of the VAR variable matrix $\mathbf{Y}$ and has the same sampling distribution as $\mathbf{Y}$. The resulting VAR lag variable matrix $\mathbf{X}$ corresponding to $\mathbf{Y}^*$ is written as $\mathbf{X}^*$. Then the problem of computing moments of $\mathbf{G}$ conditional on observation $\mathbf{Y}$ becomes the problem of computing moments of $\mathbf{X}^{*'}\mathbf{X}^*$ conditional on $(\mathbf{\Phi}, \mathbf{\Sigma})$ and $\mathbf{Y}$, e.g.,

$$\begin{aligned} \mathbb{E}(\mathbf{G} \mid \mathbf{Y}) &= \mathbb{E}\{\mathbb{E}_{(\mathbf{X}^*|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}^{*'}\mathbf{X}^*) \mid \mathbf{Y}\}, \\ \mathbb{E}\{\mathbf{G\Phi} \mid \mathbf{Y}\} &= \mathbb{E}\{\mathbb{E}_{(\mathbf{X}^*|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}^{*'}\mathbf{X}^*)\mathbf{\Phi} \mid \mathbf{Y}\}. \end{aligned}$$

A general approach of computing such posterior quantities is given in the next section.

3 General Algorithms Using Generated Data as Latent Parameters

Suppose that observed data $\mathbf{X}$ for given unknown parameters $\boldsymbol{\theta}$ has density $f(\mathbf{x}|\boldsymbol{\theta})$. The prior employed by the researcher is $\pi(\boldsymbol{\theta})$ (which may be informative or noninformtive). Let $\mathbf{X}^*$ be a random vector (or a matrix) with density $f^*(\mathbf{x}^* \mid \boldsymbol{\theta})$. Let $h(\boldsymbol{\theta})$ be a function of the parameters $\boldsymbol{\theta}$. In practice $\mathbf{X}^*$ may have the same sampling distribution as the data $\mathbf{X}$, which is the case for the VAR model in this paper. We are interested in the posterior mean of $[\mathbb{E}_{(\mathbf{X}^*|\boldsymbol{\theta})}\{g(\mathbf{X}^*)\}]h(\boldsymbol{\theta})$ given data $\mathbf{X}$.

The foundation of our algorithm is the following fact.

$$\mathbb{E}(\mathbb{E}_{(\mathbf{X}^*|\boldsymbol{\theta})}\{g(\mathbf{X}^*)\}h(\boldsymbol{\theta}) \mid \mathbf{X}) = \frac{\int\Big\{\int g(\mathbf{x}^*)f^*(\mathbf{x}^* \mid \boldsymbol{\theta})d\mathbf{x}^*\Big\}h(\boldsymbol{\theta})f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}$$

$$
\begin{aligned}
&= \frac{\int\Big\{\int g(\mathbf{x}^*)h(\boldsymbol{\theta})f^*(\mathbf{x}^* \mid \boldsymbol{\theta})d\mathbf{x}^*\Big\}f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})d\boldsymbol{\theta}}{\int\Big\{\int f^*(\tilde{\mathbf{x}}^* \mid \tilde{\boldsymbol{\theta}})d\tilde{\mathbf{x}}^*\Big\}f(\mathbf{X} \mid \tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\boldsymbol{\theta}}} \\
&= \int\int \{g(\mathbf{x}^*)h(\boldsymbol{\theta})\}\,\pi(\mathbf{x}^*, \boldsymbol{\theta} \mid \mathbf{X})\,d\mathbf{x}^* d\boldsymbol{\theta},
\end{aligned}
$$

where

$$
\pi^*(\mathbf{x}^*, \boldsymbol{\theta} \mid \mathbf{X}) = \frac{f^*(\mathbf{x}^* \mid \boldsymbol{\theta})f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})}{\int\int f^*(\tilde{\mathbf{x}}^* \mid \tilde{\boldsymbol{\theta}})f(\mathbf{X} \mid \tilde{\boldsymbol{\theta}})\pi(\tilde{\boldsymbol{\theta}})d\tilde{\mathbf{x}}^* d\tilde{\boldsymbol{\theta}}}. \tag{21}
$$

The asterisk '*' reflects the fact that the density involves simulated data $\mathbf{X}^*$. The above equations indicate that the mean of function $[\mathbb{E}_{(\mathbf{X}^*|\boldsymbol{\theta})}\{g(\mathbf{X}^*)\}]h(\boldsymbol{\theta})$ under the posterior of $\boldsymbol{\theta}$ given data $\mathbf{X}$ is the mean of $g(\mathbf{X}^*)h(\boldsymbol{\theta})$ under the posterior of $(\mathbf{X}^*, \boldsymbol{\theta})$ given data $\mathbf{X}$. Our task becomes simulating the joint posterior distribution of $(\mathbf{X}^*, \boldsymbol{\theta})$ given data $\mathbf{X}$. An MCMC method can be used for this purpose. Our approach is made computationally feasible by the law of iteration of conditional expectations. If we have a random sample $(\mathbf{X}_k^*, \boldsymbol{\theta}_k)$, $k = 1, \cdots, M$ from the joint distribution (21), we can estimate $\mathbb{E}(\mathbb{E}_{(\mathbf{X}^*|\boldsymbol{\theta})}\{g(\mathbf{X}^*)\}h(\boldsymbol{\theta}) \mid \mathbf{X})$ using the result

$$
\widehat{\mathbb{E}}(\mathbb{E}_{(\mathbf{X}^*|\boldsymbol{\theta})}\{g(\mathbf{X}^*)\}h(\boldsymbol{\theta}) \mid \mathbf{X}) = \widehat{\mathbb{E}}_{((\mathbf{X}^*, \boldsymbol{\theta})|\mathbf{X})}\{g(\mathbf{X}^*)h(\boldsymbol{\theta})\} = \frac{1}{M}\sum_{k=1}^{M} g\{\mathbf{X}_k^*\}h(\boldsymbol{\theta}_k).
$$

We propose two options to run MCMC simulations to generate $(\mathbf{X}^*, \boldsymbol{\theta})$.

Option 1: We will sample from $\pi^*(\boldsymbol{\theta} \mid \mathbf{X}^*, \mathbf{X}) \propto f^*(\mathbf{X}^* \mid \boldsymbol{\theta})f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$, the distribution of parameter $\boldsymbol{\theta}$ conditional on both the observation $\mathbf{X}$ and simulated data $\mathbf{X}^*$.

Suppose that at the beginning of cycle k we have sampled $(\mathbf{X}_{k-1}^*, \boldsymbol{\theta}_{k-1}^*)$.

Step 1. Simulate $\mathbf{X}_k^* \sim f^*(\mathbf{x}^* \mid \boldsymbol{\theta}_{k-1})$.

Step 2. Simulate $\boldsymbol{\theta}_k \sim \pi^*(\boldsymbol{\theta} \mid \mathbf{X}_k^*, \mathbf{X}) \propto f^*(\mathbf{X}_k^* \mid \boldsymbol{\theta})f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Option 2: If it is costly to simulate from the full conditional distribution $\pi^*(\boldsymbol{\theta} \mid \mathbf{X}^*, \mathbf{X})$, we may simulate $\boldsymbol{\theta}$ from the partial conditional distribution $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ and then add a Metropolis-Hastings step to the MCMC algorithm.

Step 1. Simulate $\mathbf{X}_k^* \sim f^*(\mathbf{x}^* \mid \boldsymbol{\theta}_{k-1})$.

Step 2. Simulate $\tilde{\boldsymbol{\theta}} \sim \pi(\boldsymbol{\theta} \mid \mathbf{X}) \propto f(\mathbf{X} \mid \boldsymbol{\theta})\pi(\boldsymbol{\theta})$.

Step 3. Simulate $u \sim$ uniform $(0, 1)$.

Step 4. Let

$$\boldsymbol{\theta}_k \;=\; \begin{cases} \tilde{\boldsymbol{\theta}}, & \text{if } u \le \alpha(\boldsymbol{\theta}_{k-1}, \tilde{\boldsymbol{\theta}}, \mathbf{X}_k^*), \\ \boldsymbol{\theta}_{k-1}, & \text{otherwise.} \end{cases}$$

where

$$\begin{aligned} \alpha(\boldsymbol{\theta}_{k-1}, \tilde{\boldsymbol{\theta}}, \mathbf{X}_k^*) &= \min\left(1, \frac{\pi^*(\tilde{\boldsymbol{\theta}} \mid \mathbf{X}_k^*, \mathbf{X})}{\pi^*(\boldsymbol{\theta}_{k-1} \mid \mathbf{X}_k^*, \mathbf{X})} \frac{\pi(\boldsymbol{\theta}_{k-1} \mid \mathbf{X})}{\pi(\tilde{\boldsymbol{\theta}} \mid \mathbf{X})}\right) \\ &= \min\left(1, \frac{f^*(\mathbf{X}_k^* \mid \tilde{\boldsymbol{\theta}})}{f^*(\mathbf{X}_k^* \mid \boldsymbol{\theta}_{k-1})}\right). \end{aligned}$$

In the Metropolis-Hastings step $\pi^*(\boldsymbol{\theta} \mid \mathbf{X}_k^*, \mathbf{X})$ is the target density, $\pi(\boldsymbol{\theta} \mid \mathbf{X})$ is the candidate-generating density that happens to be independent of the current state (i.e., $\tilde{\boldsymbol{\theta}}$ generated in Step 2 of the algorithm does not depend on $\boldsymbol{\theta}_{k-1}$). Tierney (1994) calls this Markov Chain an 'independence chain'. For a lucid illustration of Metropolis-Hastings algorithms, see Chib and Greenberg (1995).

In general, the two options may incur different costs in terms of computer programming and CPU time but they should produce the same posterior if the MCMC cycles are long enough. In the following we make more explicit the MCMC algorithms for the VAR model based on the two options and compare the results.

4 Bayesian Computation of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ in the VAR Model

4.1 Algorithms for simulating posteriors of $(\boldsymbol{\Phi}, \boldsymbol{\Sigma})$ in the VAR model

We use MCMC methods to sample from the posterior. In particular, we use the Gibbs sampling method illustrated by Gelfand and Smith (1990). We denote the latent variables as $\mathbf{X}^*$ and $\mathbf{Y}^*$.

Fact 2 *Consider the constant-Jeffreys prior (17). Suppose that at cycle k we have simulated $(\boldsymbol{Y}_{k-1}^*, \boldsymbol{\Phi}_{k-1}, \boldsymbol{\Sigma}_{k-1})$.*

(a) The algorithm in Option 1 is as follows.

Step 1. Simulate $\boldsymbol{y}_{k,t}^* \sim MVN(\boldsymbol{c} + \sum_{i=1}^{L} \boldsymbol{y}_{k,t-i}^* \boldsymbol{B}_{k-1,i},\, \boldsymbol{\Sigma}_{k-1})$, for $t = 1, \cdots, T$. Define

$$\boldsymbol{Y}_k^* \;=\; \begin{pmatrix} \boldsymbol{y}_{k,1}^* \\ \vdots \\ \boldsymbol{y}_{k,T}^* \end{pmatrix} \;\; \textit{and} \;\; \boldsymbol{X}_k^* \;=\; \begin{pmatrix} 1 & \boldsymbol{y}_{k,-1}^* & \cdots & \boldsymbol{y}_{k,-L}^* \\ 1 & \boldsymbol{y}_{k,0}^* & \cdots & \boldsymbol{y}_{k,1-L}^* \\ \cdot & \cdot & \cdots & \cdot \\ 1 & \boldsymbol{y}_{k,T-1}^* & \cdots & \boldsymbol{y}_{k,T-L}^* \end{pmatrix}.$$

Step 2. Simulate $\boldsymbol{\phi}_k = vec(\boldsymbol{\Phi}_k) \sim MVN(\widehat{\boldsymbol{\phi}}_k,\ \boldsymbol{\Sigma}_{k-1} \otimes (\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_k^{*\prime}\boldsymbol{X}_k^*)^{-1})$, *where*

$$\widehat{\boldsymbol{\phi}}_k \quad = \quad vec\Big\{\Big(\boldsymbol{X}'\boldsymbol{X} + \boldsymbol{X}_k^{*\prime}\boldsymbol{X}_k^*\Big)^{-1}\Big(\boldsymbol{X}'\boldsymbol{Y} + \boldsymbol{X}_k^{*\prime}\boldsymbol{Y}_k^*\Big)\Big\}. \tag{22}$$

Step 3. Simulate $\boldsymbol{\Sigma}_k \sim$ *Inverse Wishart* $(\boldsymbol{S}_k, 2T)$, *where*

$$\boldsymbol{S}_k \quad = \quad (\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Phi}_k)'(\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\Phi}_k) + (\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\boldsymbol{\Phi}_k)'(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\boldsymbol{\Phi}_k).$$

(b) The algorithm in Option 2 is as follows.

Step 1. Do the same as Step 1 in Part (a).

Step 2. Simulate $\widetilde{\boldsymbol{\phi}} \equiv vec(\widetilde{\boldsymbol{\Phi}}) \sim MVN(\widehat{\boldsymbol{\phi}}_{MLE},\ \widetilde{\boldsymbol{\Sigma}}_{k-1} \otimes (\boldsymbol{X}'\boldsymbol{X})^{-1})$, *where*

$$\widehat{\boldsymbol{\phi}}_{MLE} \quad = \quad vec\{(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{Y})\}. \tag{23}$$

Step 3. Simulate $\widetilde{\boldsymbol{\Sigma}} \sim$ *Inverse Wishart* $(\boldsymbol{S}_{0,k}, T)$, *where*

$$\boldsymbol{S}_{0,k} \quad = \quad (\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\Phi}})'(\boldsymbol{Y} - \boldsymbol{X}\widetilde{\boldsymbol{\Phi}}).$$

Step 4. Simulate $u \sim\ uniform(0,1)$.

Step 5. Define

$$(\boldsymbol{\phi}_k, \boldsymbol{\Sigma}_k) \quad = \quad \begin{cases} (\widetilde{\boldsymbol{\phi}}, \widetilde{\boldsymbol{\Sigma}}), & \text{if } u \le \alpha_k, \\ (\boldsymbol{\phi}_{k-1}, \boldsymbol{\Sigma}_{k-1}), & \text{otherwise,} \end{cases}$$

where

$$\alpha_k = \min\left(1, \frac{|\widetilde{\boldsymbol{\Sigma}}|^{-T/2} etr\left\{-\frac{1}{2}(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\widetilde{\boldsymbol{\Phi}})\widetilde{\boldsymbol{\Sigma}}^{-1}(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\widetilde{\boldsymbol{\Phi}})'\right\}}{|\boldsymbol{\Sigma}_{k-1}|^{-T/2} etr\left\{-\frac{1}{2}(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\boldsymbol{\Phi}_{k-1})\boldsymbol{\Sigma}_{k-1}^{-1}(\boldsymbol{Y}_k^* - \boldsymbol{X}_k^*\boldsymbol{\Phi}_{k-1})'\right\}}\right).$$

Fact 3 *Consider the constant-RATS prior (18). Suppose that at cycle k we have* $(\boldsymbol{Y}_{k-1}^*, \boldsymbol{\Phi}_{k-1}, \boldsymbol{\Sigma}_{k-1})$.
(a) The algorithm in Option 1 is the same as Part (a) of Fact 2, except the degree freedom for the Inverse Wishart distribution in Step 3 is $2T + Lp$ *instead of* $2T$.
(b) The algorithm in Option 2 is the same as Part (b) of Fact 2, except the degree freedom for the Inverse Wishart distribution in Step 3 is $T + Lp$ *instead of* T.

4.2 Computing the expected posterior loss

The Bayesian risk under the entropy loss can be computed using the posterior distribution generated by the MCMC procedure. Given the estimate $(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}})$, which is computed for a given data sample $\mathbf{Y}$, the expected posterior loss is $\mathbb{E}L_E(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}, \mathbf{\Phi}, \mathbf{\Sigma}|\mathbf{Y}) = \mathbb{E}[L_{E1}(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}|\mathbf{Y}) + L_{E2}(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}|\mathbf{Y})]$.

Recall that

$$\begin{aligned} L_{E1}(\widehat{\mathbf{\Sigma}}; \mathbf{\Sigma}) &= \frac{T}{2}\{tr(\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}) - \log|\widehat{\mathbf{\Sigma}}^{-1}\mathbf{\Sigma}| - p\}, \\ L_{E2}(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}; \mathbf{\Phi}, \mathbf{\Sigma}) &= \frac{1}{2}tr[\widehat{\mathbf{\Sigma}}^{-1}\{(\mathbf{\Phi} - \hat{\mathbf{\Phi}})'\mathbf{G}(\mathbf{\Phi} - \hat{\mathbf{\Phi}})\}, \end{aligned}$$

and $\mathbf{G} = \mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$.

In the kth MCMC cycle (k=1, 2, ..., M) $\mathbf{\Sigma}_k$ can be decomposed as $\mathbf{\Sigma}_k = \mathbf{Q}_k\mathbf{D}_k\mathbf{Q}_k'$, where $\mathbf{D}_k = diag(\mathbf{d}_{k1}, \mathbf{d}_{k2}, .., \mathbf{d}_{kp})$ is the diagonal matrix consisting of eigenvalues of $\mathbf{\Sigma}_k$, and $\mathbf{Q}_k$ is an orthogonal matrix with $\mathbf{Q}_k\mathbf{Q}_k' = I$.

The expected posterior loss of the L_E can be computed as the sum of the two parts,

$$\widehat{\mathbb{E}}L_{E1}(\widehat{\mathbf{\Sigma}}, \mathbf{\Sigma}|\mathbf{Y}) = \frac{T}{2}[tr(\widehat{\mathbf{\Sigma}}^{-1}\frac{1}{M}\sum_{k=1}^{M}\mathbf{\Sigma}_k) + \log|\widehat{\mathbf{\Sigma}}| - p - \frac{1}{M}\sum_{k=1}^{M}\sum_{i=1}^{p} log|\mathbf{d}_{ki}|],$$

and

$$\begin{aligned} \widehat{\mathbb{E}}L_{E2}(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}, \mathbf{\Phi}, \mathbf{\Sigma}|\mathbf{Y}) = \frac{1}{2}[tr(\widehat{\mathbf{\Sigma}}^{-1}\frac{1}{M}\sum_{k=1}^{M}\mathbf{\Phi}_k'\mathbf{X}_k^{*\prime}\mathbf{X}_k^*\mathbf{\Phi}_k) + tr(\widehat{\mathbf{\Phi}}\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Phi}}'\frac{1}{M}\sum_{k=1}^{M}\mathbf{X}_k^{*\prime}\mathbf{X}_k^*)] \\ -tr(\widehat{\mathbf{\Sigma}}^{-1}\widehat{\mathbf{\Phi}}'\frac{1}{M}\sum_{k=1}^{M}\mathbf{X}_k^{*\prime}\mathbf{X}_k^*\mathbf{\Phi}_k). \end{aligned}$$

Note that both parts are functions of moments of simulated $\mathbf{\Sigma}$, $\mathbf{\Phi}$, and $\mathbf{X}^*$ in the MCMC procedure. The moments of the simulated parameters can be computed in the MCMC cycles, just as the posterior mean, without the need of storing the parameters simulated in all MCMC cylces. Therefore the average posterior loss is computed at little additional cost in the MCMC simulations.

5 MCMC Simulation Results Without Using the Closed-Form Expression of $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$

In this section for a generated data sample posterior means and the Bayesian estimates are computed under the constant-RATS prior using alternative MCMC algorithms. The first algorithm is the algorithm Option 1 in Section 4 that makes use of full conditional posteriors. The second algorithm

is Option 2 of Section 4 which is a Metropolized MCMC drawing from partial conditional posteriors. These two algorithms should produce the same result if the number of the Markov Chain is set long enough. The number of MCMC cycles is set at 10000 in our simulations.

We consider the following example:

Example 1 The true parameters are

$$\mathbf{\Sigma} = \begin{pmatrix} 0.5 & 0 & 0 & 0 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 1.5 & 0 & 0 \\ 0 & 0 & 0 & 2.0 & 0 \\ 0 & 0 & 0 & 0 & 2.5 \end{pmatrix}, \quad \mathbf{\Phi} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 \\ 0.5 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0 & 0 \\ 0 & 0 & 0.5 & 0 & 0 \\ 0 & 0 & 0 & 0.5 & 0 \\ 0 & 0 & 0 & 0 & 0.5 \end{pmatrix}.$$

We generate a data set of 100 obsrevations using a VAR with the above true parameters. The MLE for the data sample is

$$\widehat{\mathbf{\Sigma}}_{MLE} = \begin{pmatrix} 0.4703 & 0.0620 & -0.0541 & 0.1348 & -0.0488 \\ 0.0620 & 0.8802 & 0.0084 & -0.0851 & 0.1356 \\ -0.0541 & 0.0084 & 1.2911 & 0.0047 & -0.0829 \\ 0.1348 & -0.0851 & 0.0047 & 1.7684 & -0.4500 \\ -0.0488 & 0.1356 & -0.0829 & -0.4500 & 2.0876 \end{pmatrix},$$

$$\widehat{\mathbf{\Phi}}_{MLE} = \begin{pmatrix} -0.0451 & 0.1045 & 0.0111 & 0.0983 & 0.2011 \\ 0.5258 & 0.2447 & 0.1905 & -0.1246 & -0.2422 \\ 0.0246 & 0.4174 & -0.0321 & 0.0789 & 0.2711 \\ -0.0403 & -0.0336 & 0.5014 & -0.0028 & -0.1279 \\ -0.0471 & 0.0154 & 0.0940 & 0.3681 & 0.0045 \\ -0.0139 & -0.0029 & -0.0723 & 0.0595 & 0.3507 \end{pmatrix}.$$

The mean of posterior simulated by using MCMC algorithm Option 1 is

$$\widehat{\mathbf{\Sigma}}_{Mean} = \begin{pmatrix} 0.5059 & 0.0647 & -0.0573 & 0.1455 & -0.0520 \\ 0.0647 & 0.9437 & 0.0073 & -0.0885 & 0.1486 \\ -0.0573 & 0.0073 & 1.3872 & 0.0025 & -0.0890 \\ 0.1455 & -0.0885 & 0.0025 & 1.9090 & -0.4943 \\ -0.0520 & 0.1486 & -0.0890 & -0.4943 & 2.2468 \end{pmatrix},$$

$$\widehat{\mathbf{\Phi}}_{Mean} = \begin{pmatrix} -0.0464 & 0.1016 & 0.0129 & 0.0961 & 0.2012 \\ 0.5245 & 0.2442 & 0.1889 & -0.1307 & -0.2434 \\ 0.0269 & 0.4180 & -0.0315 & 0.0774 & 0.2759 \\ -0.039 & -0.0324 & 0.5026 & -0.0016 & -0.1303 \\ -0.0459 & 0.0166 & 0.0932 & 0.3703 & 0.0036 \\ -0.0128 & -0.0007 & -0.0732 & 0.0610 & 0.3452 \end{pmatrix}.$$

The Bayesian estimate of $(\mathbf{\Sigma}, \mathbf{\Phi})$ simulated using Option 1 is

$$\widehat{\mathbf{\Sigma}}_{E} = \begin{pmatrix} 0.5414 & 0.0690 & -0.0607 & 0.1552 & -0.0562 \\ 0.0690 & 1.0099 & 0.0078 & -0.0953 & 0.1599 \\ -0.0607 & 0.0078 & 1.4840 & 0.0022 & -0.0965 \\ 0.1552 & -0.0953 & 0.0022 & 2.0431 & -0.5283 \\ -0.0562 & 0.1599 & -0.0965 & -0.5283 & 2.4018 \end{pmatrix},$$

Table 2: Posterior Risks of the Estimates in Example 1

	L_{E1}	L_{E2}	L_E
$\widehat{\mathbf{\Sigma}}_{Mean}, \widehat{\mathbf{\Phi}}_{Mean}$	7.9504	26.1078	34.0582
$\widehat{\mathbf{\Sigma}}_E, \widehat{\mathbf{\Phi}}_E$	8.5083	16.3292	24.8375

$$\widehat{\mathbf{\Phi}}_E = \begin{pmatrix} \text{-0.0405} & 0.0918 & 0.0153 & 0.0836 & 0.1826 \\ 0.5721 & 0.2391 & 0.1815 & \text{-0.1353} & \text{-0.2418} \\ 0.0313 & 0.4644 & \text{-0.0360} & 0.0780 & 0.2729 \\ \text{-0.035} & \text{-0.0327} & 0.5478 & 0.0057 & \text{-0.1331} \\ \text{-0.047} & 0.0200 & 0.0916 & 0.4165 & 0.0063 \\ \text{-0.0143} & 0.0045 & \text{-0.0742} & 0.0595 & 0.3890 \end{pmatrix}.$$

We find that the algorithms Options 1 and 2 produce estimates that are quite similar (see the appendix for the estimates obtained using Option 2). This is not surprising in light of the fact that the acceptance rate at the Metropolis-Hastings step in Option 2 is about 94%. As is reported earlier, the MCMC procedure using closed-form expression of $\mathbf{G}$ breaks down over forty percent of the time for the set of parameters used in the example, but is applicable when the diagonal elements of $\mathbf{\Phi}$ is 0.2. We also experimented with the parameterization that allows for simulation of posteriors under three alternative algorithms. The first and second are the MCMC procedures Options 1 and 2. The third algorithm is the one discussed in Section 2.3, in which the frequentist expectation $\mathbb{E}_{(\mathbf{Y}|\mathbf{\Phi},\mathbf{\Sigma})}(\mathbf{X}'\mathbf{X})$ is computed using the closed-form expression in Lemma 2. For MCMC cycles of length 10,000 the three procedures produce results that are close to be identical. Note the standard procedure that uses closed-form expression for $\mathbf{G}$ does not simulate data as latent parameters, while Options 1 and 2 add hundreds of latent parameters. However, following Options 1 and 2 after integrating out these latent parameters the simulated posteriors for $(\mathbf{\Sigma}, \mathbf{\Phi})$ are almost exactly the same as that produced by the standard algorithm. Given the large number of parameters in $\mathbf{\Sigma}$ and $\mathbf{\Phi}$ (45) and the number of data generated as latent parameters (500), the MCMC algorithms of Options 1 and 2 are surprisingly efficient. Both algorithms take about one minute to finish the entire computation on a 1.7 GHz Pentium 4 machine. There is little difference in computation time using Options 1 and 2.

For the sample generated in Example 1, the posterior risk of the posterior mean estimator (34.0582) is about 37% larger than that of the Bayesian estimator (24.8375). The lower overall posterior risk of the Bayesian estimator is achieved by substantially lowering the risk of the quadratic part, from 26.1078 for the posterior mean to 16.3292. The first term of the loss related to $\mathbf{\Sigma}$ under

Table 3: Loss of the estimates in Example 1

	L_{E1}	L_{E2}	L_E
$\widehat{\mathbf{\Sigma}}_{MLE}, \widehat{\mathbf{\Phi}}_{MLE}$	0.1953	0.3566	0.5519
$\widehat{\mathbf{\Sigma}}_{Mean}, \widehat{\mathbf{\Phi}}_{Mean}$	0.1432	0.3353	0.4786
$\widehat{\mathbf{\Sigma}}_{E}, \widehat{\mathbf{\Phi}}_{E}$	0.1195	0.2730	0.3925

the Bayesian estimator is slightly larger compared to that under the posterior mean estimator. As is noted earlier, the Bayesian estimator improves L_{E2}-related risk with a tradeoff of larger L_{E1}-related risk. Table 2 shows that the Bayesian estimator induces lower posterior risk than the posterior mean by making the L_{E2}-related risk substantially lower and the L_{E1}-related risk only slightly higher.

To compare the alternative estimates we calculate the value of the entropy loss $L_E(\widehat{\mathbf{\Phi}}, \widehat{\mathbf{\Sigma}}, \mathbf{\Phi}, \mathbf{\Sigma})$ with $(\mathbf{\Phi}, \mathbf{\Sigma})$ being the true parameters. In terms of the loss of the estimates evaluated at the true parameters, the Bayesian estimates for the data sample are better than the posterior mean and the MLE. It is well known that finite sample MLE of the variance components of $\mathbf{\Sigma}$ has a downward bias. For this sample the posterior mean under the constant-RATS prior partially corrects the bias of the MLE. Theorem 1 shows that the Bayesian estimates for the variance components of $\mathbf{\Sigma}$ are larger than those for the posterior mean. For this sample most of the variance elements in the posterior mean $\mathbf{\Sigma}$ happen to be smaller than the true parameters. It is not surprising that the Bayesian estimator incurs smaller L_{E1}-related loss at the point of true parameters than the posterior mean although the latter produces smaller average L_{E1}-related posterior loss.

The MLE of $\mathbf{\Phi}$ tends to have a downward bias if the true VAR(1) parameters represent positive auto-correlations. Applying the constant prior on $\mathbf{\Phi}$ does not make the posterior mean of $\mathbf{\Phi}$ deviate much from the MLE, hence the posterior mean also shows a downward bias. The discussion in Section 2 suggests that the Bayesian estimate for $\mathbf{\Phi}$ is likely to be larger than the posterior mean. The conjecture turns out to be true for the estimates of $\mathbf{\Phi}$.

6 Estimating a VAR Using Macroeconomic Data of the U.S.

In the past two decades, VAR models have been widely used for analyzing multivariate time series macroeconomic data. In the following, we compare Bayesian estimates and the posterior mean estimates of a six-variable VAR using quarterly data of the U.S. economy from 1959Q1 to 2001Q4. The lag length of the VAR is one.[4] The variables are real GDP, GDP deflator, world commodity price, M2 money stock, non-borrowed reserves, and the federal funds rate. The commodity price data are obtained from the International Monetary Fund, 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 in logarithms. These variables frequently appear in macroeconomics related VARs (e.g., Sims 1992, Gordon and Leeper 1994, Sims and Zha 1998, and Christiano et al. 1999).

The posterior mean and Bayesian estimator of $(\mathbf{\Sigma}, \mathbf{\Phi})$ (computed via Option 1) are reported in the following.

$$\widehat{\mathbf{\Sigma}}_{Mean} = \begin{pmatrix} 0.00442 & 0.00181 & 0.00164 & 0.00329 & 0.00075 & 0.00005 \\ 0.00181 & 0.00076 & 0.00069 & 0.00136 & 0.00035 & 0.00002 \\ 0.00164 & 0.00069 & 0.00239 & 0.00120 & 0.00022 & 0.00011 \\ 0.00329 & 0.00136 & 0.00120 & 0.00250 & 0.00061 & 0.00002 \\ 0.00075 & 0.00035 & 0.00022 & 0.00061 & 0.00141 & -0.00015 \\ 0.00005 & 0.00002 & 0.00011 & 0.00002 & -0.00015 & 0.00010 \end{pmatrix},$$

$$\widehat{\mathbf{\Phi}}_{Mean} = \begin{pmatrix} 7.63542 & 3.04415 & 3.48148 & 5.59176 & 2.34007 & 0.02463 \\ -0.47129 & -0.59331 & -0.68718 & -1.07988 & -0.45012 & -0.01141 \\ -0.78591 & 0.63318 & -0.60194 & -0.57308 & -0.20369 & -0.03752 \\ -0.10012 & -0.02430 & 0.97518 & -0.07117 & -0.05809 & 0.01857 \\ 1.11321 & 0.46930 & 0.62947 & 1.82868 & 0.35210 & 0.02071 \\ 0.09832 & 0.04369 & 0.08666 & 0.04478 & 0.99211 & -0.00034 \\ 0.08643 & 0.17231 & -0.00485 & 0.06953 & 0.15719 & 0.87980 \end{pmatrix}.$$

$$\widehat{\mathbf{\Sigma}}_{E} = \begin{pmatrix} 0.00554 & 0.00228 & 0.00216 & 0.00413 & 0.00107 & 0.00006 \\ 0.00228 & 0.00096 & 0.00092 & 0.00171 & 0.00048 & 0.00002 \\ 0.00216 & 0.00092 & 0.00328 & 0.00157 & 0.00038 & 0.00017 \\ 0.00413 & 0.00171 & 0.00157 & 0.00315 & 0.00086 & 0.00002 \\ 0.00107 & 0.00048 & 0.00038 & 0.00086 & 0.00169 & -0.00016 \\ 0.00006 & 0.00002 & 0.00017 & 0.00002 & -0.00016 & 0.00011 \end{pmatrix},$$

[4] A two-lag and a four-lag VAR produce results similar in nature to the one-lag model. We choose to report the result of the one-lag model since it takes up the least space.

Table 4: Posterior Risks of the Estimates, U.S. Macroeconomic Data 1959Q1-2001Q4

	L_{E1}	L_{E2}	L_E
$\widehat{\mathbf{\Sigma}}_{Mean}, \widehat{\mathbf{\Phi}}_{Mean}$	10.7802	6798.9789	6809.7590
$\widehat{\mathbf{\Sigma}}_E, \widehat{\mathbf{\Phi}}_E$	25.1250	103.1036	128.2286

$$\widehat{\mathbf{\Phi}}_E = \begin{pmatrix} 7.61187 & 3.03431 & 3.46824 & 5.57455 & 2.33248 & 0.02446 \\ -0.46742 & -0.59135 & -0.69031 & -1.07859 & -0.4472 & -0.00765 \\ -0.77483 & 0.64331 & -0.54087 & -0.57417 & -0.23118 & -0.02623 \\ -0.11667 & -0.03135 & 1.00848 & -0.08292 & -0.04781 & 0.00945 \\ 1.12139 & 0.46952 & 0.58962 & 1.84098 & 0.33967 & 0.01283 \\ 0.08514 & 0.03806 & 0.07128 & 0.03567 & 1.03748 & 0.00377 \\ 0.06712 & 0.16046 & -0.07539 & 0.05755 & 0.11241 & 0.98440 \end{pmatrix}.$$

Table 4 reports the posterior risks of the posterior mean and the Bayesian estimator. The Bayesian estimator dominates the posterior mean by a large margin. The large difference in the posterior risk is mainly due to the difference in the risks of the quadratic term L_{E2}, which is proportional to $\mathbf{X}'\mathbf{X}$. $\mathbf{X}'\mathbf{X}$ is quite large in this application, hence with a larger $\widehat{\mathbf{\Sigma}}$ the Bayesian estimate substantially reduces the posterior risk, compared with the posterior mean.

7 Concluding Remarks

In this paper we investigate properties of Bayesian estimators of $(\mathbf{\Phi}, \mathbf{\Sigma})$ derived from an intrinsic entropy loss function. These estimators are shown to be distinctly different from the posterior mean. The entropy loss of $(\mathbf{\Phi}, \mathbf{\Sigma})$ is non-separable in $\mathbf{\Phi}$ and $\mathbf{\Sigma}$. Using posterior mean as the estimator is equivalent of ignoring the non-separability, consequently results in larger posterior risk under the entropy loss. Computation of Bayesian estimator under the entropy loss raises a technical challenge since the weighting matrix in the loss function involves frequentist moments of VAR variables which only in very restrictive cases can be computed in closed form. We propose algorithms that use generated data as latent parameters in numerical simulation of posteriors and computation of Bayesian estimators under the entropy loss. The algorithms are shown to be quite efficient. Our MCMC simulation scheme is of interest in its own right because it may be useful for Bayesian analysis in other problems. For example, in the Bayesian VAR literature, the priors

on $\mathbf{\Phi}$ and $\mathbf{\Sigma}$ are often considered separately. In this paper we examines the Bayesian estimator under the joint non-separable loss on $(\mathbf{\Phi}, \mathbf{\Sigma})$ under separate priors. For users of Bayesian VAR, it is of interest to experiment with joint non-informative priors for $(\mathbf{\Phi}, \mathbf{\Sigma})$ under the general principles outlined in Kass and Wasserman (1996). The joint non-informative priors generally involve frequentist moments of VAR variables. The algorithms developed in this paper cleared a major obstacle in employing these joint priors in estimation of VAR models.

References

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

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.

Bernardo, J. M. (1979). Reference posterior distributions for Bayesian inference. *Journal of Royal Statistical Society Ser. B* **41**, 113-147.

Chib, S. and E. Greenberg (1995). Understanding the Metropolis-Hastings algorithms. *American Statistician* **49**, 327-335.

Chib, S. and E. Greenberg (1996). Markov Chain Monte Carlo simulation methods in econometrics. *Econometric Theory* **12**, 327-335.

Christiano, L.J., Eichenbaum, M. and Evans C. (1999). Monetary policy shocks: What have we learned and to what end? in: J.B. Taylor and M. Woodford eds., *Handbook of Macroeconomics*, Volume **1**, 65–147.

DeJong D.N., Ingram B.F., and Whiteman C.H. (2000). A Bayesian approach to dynamic macroeconomics. *Journal of Econometrics* **98**, 203-223.

Elerian O., Chib, S. and Shephard N. (2001). Likelihood inference for discretely observed nonlinear diffusions. *Econometrica* **69**, 959-993.

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. (1989). Bayesian inference in econometric models using Monte Carlo integration. *Econometrica* **57**, 1317-1339.

Geweke J. (1993). Bayesian treatment of the independent Student-t linear model. *Journal of Applied Econometrics* **8**, s19-s40.

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.

Jeffreys, H. (1961) *Probability Theory.* Oxford University Press, New York.

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

Lutkepohl H. (1993). *Introduction to Multiple Time Series (2nd edition).* Springer-Verlag, New York.

Ni S. and Sun D. (2001). A Monte Carlo study on frequentist risks of Bayesian estimators of vector-autoregressive models based on noninformative priors. *Manuscript.*

Otrok C. and Whiteman C.H. (1998). Bayesian leading indicators: measuring and predicating economic conditions in Iowa. *International Economic Review* **39**, 997-1014.

Press, J.S. (1982). *Applied Multivariate Analysis: Using Bayesian and Frequentist Methods of Inference (2nd edition).* Krieger, Florida.

Robert C.P. (1994). *The Bayesian Choice.* Springer-Verlag, New York.

Sims, C. A. (1992). Interpreting the macroeconomic time series facts: The effects of monetary policy. *European Economic Review* **38**, 975 – 1000.

Sims, C.A. and Zha T. (1998). 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.

Tanner M. and Wang W.H. (1987). The calculation of posterior distributions by data augmentation. *Journal of American Statistical Association* **82**, 528-540.

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

Tierney L. (1994). Markov chains for exploring posterior distributions (with discussions). *Annals of Statistics* **22**, 1701-1762.

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

Appendix

For the data sample in Example 1, the results using MCMC algorithm Option 2 are very close to that obtained using algorithm Option 1. The posterior mean obtained from MCMC algorithm Option 2 is

$$\widehat{\boldsymbol{\Sigma}}_{Mean} = \begin{pmatrix} 0.5049 & 0.0666 & -0.0589 & 0.1424 & -0.0493 \\ 0.0666 & 0.9476 & 0.0079 & -0.0923 & 0.1482 \\ -0.0589 & 0.0079 & 1.3878 & 0.0038 & -0.0900 \\ 0.1424 & -0.0923 & 0.0038 & 1.8985 & -0.4805 \\ -0.0493 & 0.1482 & -0.0900 & -0.4805 & 2.2389 \end{pmatrix},$$

$$\widehat{\boldsymbol{\Phi}}_{Mean} = \begin{pmatrix} -0.0444 & 0.1044 & 0.0106 & 0.0961 & 0.2021 \\ 0.5253 & 0.2453 & 0.1898 & -0.1276 & -0.2409 \\ 0.0248 & 0.4177 & -0.0344 & 0.0804 & 0.2684 \\ -0.0394 & -0.0329 & 0.5005 & -0.0028 & -0.1264 \\ -0.0469 & 0.0151 & 0.0941 & 0.3672 & 0.0059 \\ -0.0146 & -0.0021 & -0.0721 & 0.0588 & 0.3513 \end{pmatrix}.$$

The Bayesian estimate of $(\boldsymbol{\Sigma}, \boldsymbol{\Phi})$ is

$$\widehat{\boldsymbol{\Sigma}}_{E} = \begin{pmatrix} 0.5394 & 0.0711 & -0.0627 & 0.1519 & -0.0528 \\ 0.0711 & 1.0123 & 0.0089 & -0.0993 & 0.1586 \\ -0.0627 & 0.0089 & 1.4823 & 0.0050 & -0.0963 \\ 0.1519 & -0.0993 & 0.0050 & 2.0289 & -0.5120 \\ -0.0528 & 0.1586 & -0.0963 & -0.5120 & 2.3935 \end{pmatrix},$$

$$\widehat{\boldsymbol{\Phi}}_{E} = \begin{pmatrix} -0.0390 & 0.0959 & 0.0121 & 0.0868 & 0.1844 \\ 0.5708 & 0.2409 & 0.1816 & -0.1308 & -0.2389 \\ 0.0281 & 0.4621 & -0.0370 & 0.0801 & 0.2653 \\ -0.0354 & -0.0331 & 0.5445 & 0.0052 & -0.1280 \\ -0.0473 & 0.0180 & 0.0940 & 0.4115 & 0.0075 \\ -0.0156 & 0.0024 & -0.0711 & 0.0568 & 0.3950 \end{pmatrix}.$$