Deviance Information Criterion for Comparing VAR Models[*]

Tao Zeng
Singapore Management University

Yong Li
Renmin University

Jun Yu
Singapore Management University

June 16, 2014

Abstract: Vector Autoregression (VAR) has been a standard empirical tool used in macroeconomics and finance. In this paper we discuss how to compare alternative VAR models after they are estimated by Bayesian MCMC methods. In particular we apply a robust version of deviance information criterion (RDIC) recently developed in Li et al. (2014b) to determine the best candidate model. RDIC is a better information criterion than the widely used deviance information criterion (DIC) when latent variables are involved in candidate models. Empirical analysis using US data shows that the optimal model selected by RDIC can be different from that by DIC.

JEL classification: C11, C12, G12

Keywords: Bayes factor, DIC; VAR models; Markov Chain Monte Carlo.

1 Introduction

In the past thirty years, Vector Autoregression (VAR) has been widely used to capture the linear interdependencies among multiple time series of macroeconomic variables and

[*] Yong Li, Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing, 1000872, P.R. China. Tao Zeng, School of Economics and Sim Kee Boon Institute for Financial Economics, Singapore Management University, 90 Stamford Road, Singapore 178903. Jun Yu, School of Economics and Lee Kong Chian School of Business. Email for Jun Yu: yujun@smu.edu.sg. URL: http://www.mysmu.edu/faculty/yujun/. Li gratefully acknowledges the financial support by the Fundamental Research Funds for the Central Universities, and the Research Funds of Renmin University of China. Yu thanks the Singapore Ministry of Education for Academic Research Fund under grant number MOE2011-T2-2-096. We acknowledge the helpful comments from a referee and the participants at a conference in honor Peter C.B. Phillips held at Southern Methodist University on November 1-3, 2013.

financial variables. Consequently, VAR models have evolved as a standard tool for evaluating the monetary policy, for predicting macroeconomic variables and financial variables, and for the impulse response analysis. In VAR, each variable has an equation explaining its evolution based on its own lags and the lags of the other variables. Unfortunately, typically economic theory is silent about the choice of the lag length and hence such a choice must be determined by data at hand. When the lag length is not small and when the system is of moderate size, there is a large number of parameters in VAR and, hence, the identification and the estimation of VAR may be a formidable task; see for example, Ni and Sun (2003).

In large cross-sections of time series, factor models, with a small number of common latent factors, have been employed to alleviate the problem of a large number of parameters; see, for example, Forni et al. (2003), Stock and Watson (2002a,b), Bernanke and Boivin (2003), Bovin and Ng (2005). Obviously, an important choice is the number of common factors. More often than not, the common factors are assumed to follow a VAR model. In this case, another important choice is the lag length. As in the case of the basic VAR models, usually economic theory does not offer guidance to choose the lag length in the factor VAR models. Once again, how many lags to be used is an important empirical question.

Frequentist's inferential approaches have proven difficult for estimating some VAR models. Consequently, Bayesian approaches have been increasingly popular for estimating VAR models; see, for example, Koop and Korobilis (2009), Otrok and Whiteman (1998), Kose et al. (2003, 2008), Bernanke et al. (2005), Bai and Wang (2012). In particular, with the advance of MCMC algorithms and expanded computing facility, Bayesian MCMC methods have been routinely used for estimating VAR.

The question we ask in this paper is the following. After a set of candidate VAR models have been estimated by MCMC, with some candidate models involving latent common factors, how should one select the optimal model? Model comparison is one of the most important statistical inferences that one has to face; see, for example, Phillips (2005, 2006). In the Bayesian literature, Bayes factor (BF) (Kass and Raftery (1995)) is arguably the most popular tool for Bayesian model comparison. However, BF is subject to some theoretical drawbacks as well as some computational limitations. For instance, it suffers from the well-known Jeffreys-Lindley paradox; see Robert (2001), Li and Yu (2012). For another example, calculation of BF requires the evaluation of marginal likelihood. For some VAR models, marginal likelihood involves high-dimensional integrations numerically. This is the case for the factor VAR models when factors are latent. Consequently, the implementation of BF entails high computational cost.

As an alternative approach to BF, a recent contribution to Bayesian model comparison is the Deviance information criterion (DIC) of Spiegelhalter et al. (2002). DIC can be understood as the Bayesian version of the Akaike information criterion (AIC, Akaike (1973)). As shown in Spiegelhalter et al. (2002), DIC is relatively simple to compute from the MCMC output, compared with BF. Moreover, it is immune to the Jeffreys-Lindley paradox.

However, as pointed out by Li et al. (2014b), the original DIC, developed by Spiegelhalter et al. (2002) and implemented in a Bayesian software WinBUGS, is not asymptotically justified for models that involve latent variables. Consequently, Li et al. (2014b) advocated the use of a robust version of DIC (RDIC) for comparing models involving latent variables. In this paper, we examine the performance of DIC and RDIC for comparing VAR models, with and without latent variables.

The paper is organized as follows. In Section 2, we introduce the VAR models and review the Bayesian MCMC methods. Section 3 reviews DIC and RDIC and shows how to calculate RDIC for the VAR models. Section 4 compares DIC and RDIC using real data. Section 5 concludes the paper. Appendix collects the derivations needed for computing RDIC.

2 Bayesian Analysis of VAR Models

In this section, we first give a simple description of the VAR models. The basic VAR(p) model is of the form

$$y_t = a_0 + \sum_{j=1}^{p} A_j y_{t-j} + \varepsilon_t, \tag{1}$$

where y_t is an $N \times 1$ vector containing T observations, ε_t an $N \times 1$ vector of errors which is i.i.d. $N(0, \Sigma)$, a_0 an $N \times 1$ vector of intercepts, A_j an $N \times N$ matrix of coefficients, and p the lag length.

The above VAR model can also be written in the matrix form

$$\boldsymbol{y} = XA + E,$$

or

$$Y = (I_M \otimes X)\,\alpha + \varepsilon,$$

where $\boldsymbol{y}$ is a $T \times N$ matrix which stacks the T observations on each dependent variable in columns next to one another, $E = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_T)'$, $x_t = \left(1, y_{t-1}', ..., y_{t-p}'\right)'$ and $X =$

$(x_1', ..., x_T')'$, a $T \times K$ matrix, $K = 1 + Np$, $A = (a_0, A_1, ..., A_p)'$, $Y = vec(\boldsymbol{y})$, $\alpha = vec(A)$, $\varepsilon = vec(E) \sim N(0, \Sigma \otimes I_T)$.

VAR may not be parsimonious since there may be a great many coefficients in it. With the typical sample size for macroeconomic variables, it is not easy to obtain reliable estimates when the dimension of the parameter space is very large. Bayesian estimation is attractive because the prior information provides a logical and formally consistent way of introducing shrinkage to reduce the over-parametrizations problem (Koop and Korobilis (2009)).

Following Koop and Korobilis (2009), for a given lag length p, we implement Bayesian analysis with the natural conjugate priors:

$$\alpha|\Sigma \sim N\left(0, \Sigma \otimes \underline{V}\right),$$

and

$$\Sigma^{-1} \sim Wishart\left(\underline{S}^{-1}, \underline{v}\right),$$

where

$$\alpha = vec\left(A\right),\ \underline{V} = 10 \times I_K,\ \underline{S}^{-1} = I_N,\ \underline{v} = N + 1.$$

As a result, the posterior distribution is

$$\alpha|\Sigma, \boldsymbol{y} \sim N\left(\widehat{\alpha}, \Sigma \otimes \left(X'X\right)^{-1}\right),$$

and

$$\Sigma^{-1}|\boldsymbol{y} \sim Wishart\left(S^{-1}, T - K - N - 1\right),$$

where $\widehat{A} = \left(X'X\right)^{-1} X'\boldsymbol{y}$, $\widehat{\alpha} = vec\left(\widehat{A}\right)$ and $S = \left(\boldsymbol{y} - X\widehat{A}\right)' \left(\boldsymbol{y} - X\widehat{A}\right)$.

To reduce the dimensionality of the parameter space in (1), one may consider introducing a smaller number of dynamic factors, f_t, so that

$$\begin{aligned} y_t &= L\boldsymbol{f}_t + e_t, & (2) \\ \boldsymbol{f}_t &= \Phi_1 \boldsymbol{f}_{t-1} + \cdots + \Phi_h \boldsymbol{f}_{t-h} + \varepsilon_t, & (3) \end{aligned}$$

where $\boldsymbol{f}_t$ is the $q \times 1$ $(q < N)$ latent dynamic factor which follows a VAR specification, L the $N \times q$ dynamic factor loading, Φ_j the $q \times q$ matrix, e_t i.i.d. $N(0, \Sigma)$, and ε_t i.i.d. $N(0, Q)$. We assume $\{e_t\}_{t=1}^T$ is independent of $\{\varepsilon_t\}_{t=1}^T$. Following Bai and Wang (2012), we assume the number of dynamic factors q does not depend on h. To achieve the identification of the factor VAR model, we set the upper $q \times q$ block of L to an identity matrix, that is

$$L = \begin{bmatrix} I_q \\ * \end{bmatrix}.$$

This identification restriction was also used in Bernanke et al. (2005).

For the basic VAR models, the likelihood function, $p(\boldsymbol{y}|\boldsymbol{\theta})$, has an analytical form, where the observed data is denoted as $\boldsymbol{y} = (y_1, \cdots, y_T)'$ and $\boldsymbol{\theta}$ contains the model parameters. Hence, the posterior distribution $p(\boldsymbol{\theta}|\boldsymbol{y})$ is easy to obtain, that is,

$$p(\boldsymbol{\theta}|\boldsymbol{y}) \propto p(\boldsymbol{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}).$$

For the factor VAR model, denote the latent factors $\boldsymbol{f} = (\boldsymbol{f}_1, \boldsymbol{f}_2, \cdots, \boldsymbol{f}_T)'$. In this case, the likelihood function involves unobserved dynamic factors, that is,

$$p(\boldsymbol{y}|\boldsymbol{\theta}) = \int p(\boldsymbol{y}, \boldsymbol{f}|\boldsymbol{\theta})d\boldsymbol{f},$$

where $p(\boldsymbol{y}, \boldsymbol{f}|\boldsymbol{\theta})$ is the so-called complete data likelihood function. Not surprisingly, the estimation of the factor VAR model is more difficult by the classical estimation procedures.

However, with the help of the data-augmentation strategy of Tanner and Wang (1987) and the MCMC techniques, Bayesian approach can easily provide the full likelihood inference of the factor VAR model. The idea is to augment the parameter space from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta}, \boldsymbol{f})$. As a result, given p, h, and q, the new posterior distribution is

$$p(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y}) \propto p(\boldsymbol{\theta}, \boldsymbol{f}, \boldsymbol{y}) = p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{f})p(\boldsymbol{f}|\boldsymbol{\theta})p(\boldsymbol{\theta}), \tag{4}$$

where

$$\begin{aligned} p(\boldsymbol{y}|\boldsymbol{\theta}, \boldsymbol{f}) &= \prod_{t=1}^{T} p(\boldsymbol{y}_t|\boldsymbol{\theta}, \boldsymbol{f}), \\ p(\boldsymbol{f}|\boldsymbol{\theta}) =& p(\boldsymbol{f}_0|\boldsymbol{\theta}) \prod_{t=1}^{T} p(\boldsymbol{f}_t|\boldsymbol{f}_{t-1}, \boldsymbol{\theta}). \end{aligned}$$

MCMC techniques may be used to obtain random samples from the posterior distribution (4). Bayesian estimates of $\boldsymbol{\theta}$ and the latent volatilities $\boldsymbol{f}$ can be obtained easily via the corresponding means of random samples. Specifically, let $\{\boldsymbol{\theta}^{(j)}, \boldsymbol{f}^{(j)}, j = 1, 2, \cdots, J\}$ be the efficient random samples generated from the joint posterior distribution $p(\boldsymbol{\theta}, \boldsymbol{f}|\boldsymbol{y})$ after discarding the burn-in samples. Then the posterior mean of $\boldsymbol{\theta}, \boldsymbol{f}$ can be obtained by

$$\hat{\boldsymbol{\theta}} = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{\theta}^{(j)}, \quad \hat{\boldsymbol{f}} = \frac{1}{J}\sum_{j=1}^{J} \boldsymbol{f}^{(j)}. \tag{5}$$

Similarly, the posterior variance of $\boldsymbol{\theta}$ can be obtained by

$$\widehat{Var}(\boldsymbol{\theta}|\boldsymbol{y}) = \frac{1}{J-1}\sum_{j=1}^{J} \left(\boldsymbol{\theta}^{(j)} - \hat{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}^{(j)} - \hat{\boldsymbol{\theta}}\right)'.$$

3 Deviance Information Criterion for VAR Models

In the Bayesian literature, there are two popular tools for model comparison. The first one is BF (Kass and Raftery, 1995), while the other is DIC (Spiegelhalter et al. (2002)). Given two candidate models M_1 and M_2, BF is given by

$$B_{12} = \frac{p(\boldsymbol{y}|M_1)}{p(\boldsymbol{y}|M_2)}, \tag{6}$$

where $p(\boldsymbol{y}|M_k), k = 1, 2$ is the marginal likelihood of M_k, and can be obtained by integrating over the support of the parameters and latent states, that is,

$$p(\boldsymbol{y}|M_k) = \int_{\Omega_k \cup \Omega_\omega} p(\boldsymbol{y}, \boldsymbol{f}\,|\boldsymbol{\theta}_k, M_k) p(\boldsymbol{\theta}_k|M_k) \mathrm{d}\boldsymbol{f}\, \mathrm{d}\boldsymbol{\theta}_k, \quad \boldsymbol{\theta}_k \in \boldsymbol{\Omega}_k, k = 1, 2.$$

However, if a vague prior is adopted, BF suffers from the Jeffreys-Lindley paradox. As pointed out by Kass and Raftery (1995), when a proper prior with a very large spread is used to represent the prior ignorance, this behavior will force the BF to favor its competitive model; see Example 1 in Li et al. (2014a) for an illustration of this problem.

Moreover, calculation of BF in general necessitates the evaluation of marginal likelihood which is a marginalization over the parameter vectors in each candidate model. When the dimension of the parameter space is large, the high-dimensional integration entails high computational cost.

DIC was recently proposed and heuristically justified by Spiegelhalter et al. (2002) based on the classical asymptotic theory.[1] It is calculated in a Bayesian software WinBUGS and widely used in the Bayesian community to compare alternative models. The basic form of DIC can be expressed as follows:

$$\mathrm{DIC} = D(\bar{\boldsymbol{\theta}}) + 2P_D = \overline{D(\boldsymbol{\theta})} + P_D, \tag{7}$$

$$D(\bar{\boldsymbol{\theta}}) = -2 \log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}}), \overline{D(\boldsymbol{\theta})} = -2 \int \log p(\boldsymbol{y}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta}, \tag{8}$$

$$P_D = -2 \int [\log p(\boldsymbol{y}|\boldsymbol{\theta}) - \log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}})] p(\boldsymbol{\theta}|\boldsymbol{y}) \mathrm{d}\boldsymbol{\theta} = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}), \tag{9}$$

where $\bar{\boldsymbol{\theta}}$ is the Bayesian estimator, $\overline{D(\boldsymbol{\theta})}$ a Bayesian measure of fit which may be better considered as a measure of 'adequacy', and P_D a Bayesian measure of model complexity. One advantage of DIC is that it is immune to Jeffreys-Lindley paradox. However, it is

[1] The rigorous asymptotical justification of DIC is given in Li et al (2014b). let $\mathbf{y}_{rep} = (\mathbf{y}_{1,rep}, \mathbf{y}_{2,rep}, \cdots, \mathbf{y}_{n,rep})$ be the independent replicate data generated by the same mechanism that gives rise to the observed data $\mathbf{y}$. Spiegelhalter, et al (2002, Page 604) proposed to choose the loss function $\mathsf{L}(\mathbf{y}_{rep}, \mathbf{y}) = -2 \log p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$. Li et al (2014b) showed that such a loss function cannot be used to justify DIC. By choosing the loss function as $\mathsf{L}(\mathbf{y}_{rep}, \mathbf{y}) = -2 \log p(\mathbf{y}_{rep}|\mathbf{y})$, under regular conditions, Li et al (2014b) asymptotically justified DIC in Theorem 3.2.

important to point out that DIC addresses how well the posterior might predict future data generated by the same mechanism that gave rise to the observed data. This is different from BF which addresses how observed data are predicted by the priors. Perhaps this difference makes DIC more suitable for comparing alternative VAR models when the primary objective of VAR is forecasting.

When the likelihood function $p(\boldsymbol{y}|\boldsymbol{\theta})$ is available in closed-form, MCMC is easy to implement and DIC is easy to calculate. For a VAR model with latent variables, to facilitate MCMC simulations, the *data augmentation* strategy is often used, as shown in Section 2. With *data augmentation*, Spiegelhalter et al. (2002) suggested the following way to compute DIC:

$$\mathrm{DIC}^* = D^*(\bar{\boldsymbol{\theta}},\overline{\boldsymbol{f}}) + 2P_D^* = \overline{D^*(\boldsymbol{\theta},\boldsymbol{f})} + P_D^*, \tag{10}$$

$$D^*(\bar{\boldsymbol{\theta}},\overline{\boldsymbol{f}}) = -2\log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}},\overline{\boldsymbol{f}}), \tag{11}$$

$$\overline{D^*(\boldsymbol{\theta},\boldsymbol{f})} = -2\int \log p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{f})p(\boldsymbol{\theta},\boldsymbol{f}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{f}, \tag{12}$$

$$P_D^* = -2\int [\log p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{f}) - \log p(\boldsymbol{y}|\bar{\boldsymbol{\theta}},\overline{\boldsymbol{f}})]p(\boldsymbol{\theta},\boldsymbol{f}|\boldsymbol{y})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{f}, \tag{13}$$

where $\overline{\boldsymbol{f}}$ the Bayesian estimator of $\boldsymbol{f}$. Obviously, the latent variables $\boldsymbol{f}$ are treated in the same way as $\boldsymbol{\theta}$. This treatment greatly facilitates the calculation of DIC* from an MCMC output when the new likelihood function, $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{f})$, is available in closed-form, a case for the factor VAR model.

While computationally tractable, unfortunately, *data augmentation* invalidates the theoretical underpinnings of DIC* since it makes the likelihood function, $p(\boldsymbol{y}|\boldsymbol{\theta},\boldsymbol{f})$, non-regular. This is because the dimension of the parameter space $(\boldsymbol{\theta},\boldsymbol{f})$ increases as the sample size increases. In particular, the non-regular likelihood problem does not lead to the asymptotic normality in the posterior distribution of $(\boldsymbol{\theta},\boldsymbol{f})$. Moreover, DIC* can be highly sensitive to nonlinear transformations of latent variables and distributional representations of model specification. In a recent study, Chan and Grant (2014) showed that DIC* has a larger variability than DIC in the context of latent variable models.

To overcome this problem, Li et al. (2014b) proposed a robust version of DIC (RDIC) for comparing latent variable models. RDIC is easy to compute and robust to nonlinear transformations of latent variables and distributional representations of model specification. The RDIC is given by:

$$\mathrm{RDIC} = D(\bar{\boldsymbol{\theta}}) + 2\mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\} = D(\bar{\boldsymbol{\theta}}) + 2P_{RD}, \tag{14}$$

where

$$P_{RD} = \mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}, \tag{15}$$

with **tr** denoting the trace of a matrix,

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2 \log p(\boldsymbol{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'}, V(\bar{\boldsymbol{\theta}}) = E\left[\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\right)' |\boldsymbol{y}\right].$$

Under regular conditions, Li et al. (2014b) showed that $P_{RD} = P_D + o_p(1)$, RDIC=DIC+$o_p(1)$, where DIC, RDIC, P_D and P_{RD} are defined in (7), (14), (9), and (15), respectively. The conditions include the regular conditions to the Bayesian large sample theory; see for example, Chen (1985). In addition, it is required that the data generating process is stationary and that the model is regular so that the standard maximum likelihood theory can be applied.

We should point out that RDIC is different from Takeuchi Information Criterion (TIC, Takeuchi (1976)) which is defined as

$$\text{TIC} = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + 2\mathbf{tr}\left\{\mathbf{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\right\}$$

where $\boldsymbol{y}^{\boldsymbol{t-1}} = (y_1, y_2, \cdots, y_{t-1})$,$\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$ is the ML estimator of $\boldsymbol{\theta}$, $\mathbf{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})$ is a consistent estimate of the long run variance. TIC can be equivalently written as

$$\text{TIC} = -2\log p(\boldsymbol{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + 2\mathbf{tr}\left\{\mathbf{I}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\widehat{\Sigma}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\right\},$$

where

$$\widehat{\Sigma}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) = \mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\mathbf{J}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}),$$

is the so-called sandwich covariance matrix for $\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}$. Although $\mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}$ and $\mathbf{tr}\left\{\mathbf{I}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\widehat{\Sigma}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\right\}$ are similar, a closer comparison of them shows important differences. First, $\mathbf{tr}\left\{\mathbf{I}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\widehat{\Sigma}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})\right\}$ in TIC is based on the ML estimation, whereas $\mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}$ in RDIC is based on the Bayesian estimation. Second, $\widehat{\Sigma}(\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}})$ in TIC requires the inversion of the matrix $I(\boldsymbol{\theta})$. When the dimension of $\boldsymbol{\theta}$ is high, $I(\boldsymbol{\theta})$ may be difficult to invert. However, in RDIC, there is no need to invert any matrix. Third, like AIC, RDIC requires the model be correctly specified, but TIC relaxes this assumption.

It should be pointed out that, like AIC, RDIC does not have the consistent property for the true model. This is different from the Bayesian information criterion such as BIC (Schwarz, 1978) that approximates BF in large sample (Kass and Raftery, 1995). Compared to BF, hence, we expect RDIC tends to choose less parsimonious models. This comparison is similar to between of AIC and BIC.

The RDIC clearly requires the evaluation of the observed information matrix and the second derivative of the observed-data likelihood function. For most latent variable models, the observed-data likelihood function does not have a closed-from expression so that the second derivatives are difficult to evaluate. However, with the help of the EM algorithm, the second derivatives can be easily approximated. In particular, under the mild regularity conditions, Louis (1982) derived the observed information matrix as:

$$\begin{aligned}\mathbf{I}(\boldsymbol{\theta}) = E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}} & \left\{-\frac{\partial^2 \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} - Var_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left\{\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right\} \\ = \quad & E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left\{-\frac{\partial^2 \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\} \\ & + E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left(\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left(\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right),\end{aligned}$$

where the expectations are taken with respect to the conditional distribution of $\boldsymbol{f}$ given $\boldsymbol{y}$ and $\boldsymbol{\theta}$. Hence, the information matrix can be approximated by:

$$\begin{aligned} & E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left\{-\frac{\partial^2 \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\} \\ \approx \quad & -\frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\partial^2 \log p(\boldsymbol{y},\boldsymbol{f}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\}, \\ & E_{\boldsymbol{f}\,|\boldsymbol{y},\boldsymbol{\theta}}\left(\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}\,|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right) \approx \frac{1}{M}\sum_{m=1}^{M}\frac{\partial \log p(\boldsymbol{y},\boldsymbol{f}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}},\end{aligned}$$

For more details, one can refer to Li et al. (2014b).

In this paper, we calculate DIC, DIC* and RDIC for the basic VAR models and the VAR models with latent variables. For the basic VAR model, all three statistics are easy to compute and $P_D = P_D^*$, DIC=DIC*. For the factor VAR models, DIC is computationally much more demanding than DIC* and RDIC. DIC* is routinely reported in the literature.

4 Empirical Study

In this section, we will determine the optimal lag length in the basic VAR model (q) and optimal lag length in the factor VAR model (h) using real data. In particular, we fit both the basic VAR models and the factor VAR models to the quarterly U.S. data on the inflation rate π_t (the annual percentage change in a chain-weighted price index), the unemployment rate u_t (the seasonally adjusted unemployment rate) and the interest rate r_t (the yield on the three month Treasury bill rate). The data have been previously used by Koop and Korobilis (2009), Koop and Potter (2011). The sample period is from 1953Q1 to 2006Q3.

We first use DIC, DIC* and RDIC to select the optimal lag in the basic VAR model. In the empirical analysis, $y_t = (\pi_t, u_t, r_t)$ and $N = 3$. The prior is set at

$$\underline{V} = 10 \times I_K,\ \underline{S}^{-1} = I_M,\ \underline{v} = N + 1.$$

For the basic VAR models, we allow nine different lag lengths, i.e., $p = 1, ..., 9$. Without latent variables in the models, DIC=DIC*.

To compute DIC, we draw $12,000$ samples from the posterior distribution and discard the first $2,000$ draws. Since both the posterior distribution and the second order derivatives have the analytical form, we can easily compute RDIC using the posterior mean of the parameters. Table 1 reports DIC, P_D, RDIC and P_{RD} for the candidate models. Several conclusions can be drawn from Table 1. First, both RDIC and DIC suggest that $p = 6$ is the best model. Second, RDIC and DIC take nearly identical values in all the candidate models. This is not surprising since there is no latent variable in the basic VAR models.

P	1	2	3	4	5	6	7	8	9
P_D	17.53	26.53	35.75	45.24	54.93	**64.78**	75.23	85.92	96.77
DIC	857.72	730.86	725.17	725.87	680.3	**672.15**	676.15	687.11	681.88
P_{RD}	17.31	26.13	35.17	44.47	53.99	**63.58**	73.91	84.64	95.01
$RDIC$	857.28	730.06	724.06	724.34	678.41	**669.76**	673.5	684.53	679.01

Table 1: Model selection results for the basic VAR models

We then fit the data to the one-factor VAR model for $h = 1, 2, 3, 4$. With the latent variables involved, DIC is no longer the same as DIC*. To select the optimal lag length, we use DIC, DIC* and RDIC. For simplicity, we assume R to be a diagonal matrix. The priors are similar to those used in Bernanke et al. (2005), namely,

$$\begin{aligned} \Sigma_{ii} &\sim IG\left(3, 0.001\right), \quad L_i \sim N\left(0, \Sigma_{ii} \times I_q\right), \\ vec\left(\Phi\right)|Q &\sim N\left(0, Q \otimes \Omega_0\right), \quad Q \sim Inverse - Wishart\left(Q_0, q+2\right), \end{aligned}$$

where Σ_{ii} is the i^{th} diagonal element of R and L_i is the i^{th} row of L $(i > q)$, $\Phi = (\Phi_1, ..., \Phi_h)$. See Bernanke et al. (2005) for the construction of Ω_0 and Q_0.

Although DIC is in general difficult to calculate for latent variable models, in this particular case, it can be obtained by using the Kalman filter. Of course, DIC is computationally much more expensive to compute than DIC* and RDIC. Using the Gibbs sampler, we sample 22,000 random observations from the corresponding posterior distributions. We discard the first 2,000 observations and keep the following 20,000 as the effective samples from the posterior distribution of the parameters.

Table 2 reports DIC, P_D, DIC*, P_D^*, RDIC and P_{RD}. Some findings arise from Table 2. First, both RDIC and DIC suggest that $h = 2$ is the best model, whereas DIC* selects $h = 3$. Second, RDIC and DIC take nearly identical values in all the candidate models. However, they take quite different values from DIC*. This difference arises because the latent variables are included into the parameter space in DIC*. As previously explained, such an expansion of the parameter space undermines the theoretical justification of DIC* and should not be used to compare alternative models. In this particular case, the use of DIC* will select too big a model although it is not clear if this is the common feature of DIC* in general. Li et al (2014b) also report evidence that DIC* ranks the same models with different representations differently.

h	1	2	3	4
P_D^*	216	243	233	238
DIC^*	2294	1966	**1941**	1942
P_{RD}	5.34	4.65	6.06	6.2
$RDIC$	2120.68	**1879.1**	1882.12	1880.1
P_D	5.8131	5.9328	6.8114	7.3254
DIC	2121.9	**1881.7**	1884	1882.3

Table 2: Model selection results for the factor VAR models

5 Conclusion

This paper uses a robust deviance information criteria (RDIC) to determine the optimal lag length in the basic VAR models and in the factor VAR models. When the latent variable is treated as parameters to facilitate Bayesian parameter estimation, the widely used DIC lacks of theoretical justification. This is because that the justification of DIC relies on the validity of the standard Bayesian asymptotic theory. In particular, when the latent variable is treated as parameters, the number of parameters increases with the number of observations, making the likelihood nonregular. In the empirical analysis, we show that in the basic VAR model where there is no latent variable, DIC and RDIC select the same optimal model. However, in the factor VAR model where the factors are latent, DIC and RDIC select the different optimal model.

6 Appendix

6.1 Appendix 1: The derivation of RDIC for the basic VAR(p) models

The log-likelihood function for VAR(p) model is:

$$\log p\left(\boldsymbol{y}|\Sigma, A\right) = -\frac{TN}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma| - \frac{1}{2}tr\left[\Sigma^{-1}\left(\boldsymbol{y}-XA\right)'\left(\boldsymbol{y}-XA\right)\right].$$

Using the matrix differentiation rules of Magnus and Neudecker (1999), the first order derivative is:

$$D_A \log p\left(\boldsymbol{y}|\Sigma, A\right) = \left[vec\left(\left(\frac{1}{2}\left(\Sigma^{-1\prime} + \Sigma^{-1}\right)\left(\boldsymbol{y} - XA\right)' X\right)'\right)\right]',$$

$$D_\Sigma \log p\left(\boldsymbol{y}|\Sigma, A\right) = \left[vec\left(\left(-\frac{T}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left(\boldsymbol{y} - XA\right)'\left(\boldsymbol{y} - XA\right)\Sigma^{-1}\right)'\right)\right],$$

and the second order derivatives are

$$D_{AA} \log p\left(\boldsymbol{y}|\Sigma, A\right) = -\frac{1}{2}\left(\left(\Sigma^{-1\prime} + \Sigma^{-1}\right) \otimes X'X\right),$$

$$D_{A\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) = -\frac{1}{2}K_{NK}\left(\left[\left(\boldsymbol{y} - XA\right)' X\right]' \otimes I_N\right)\left(K_{NN} + I_{NN}\right)\left(\Sigma^{-1} \otimes \Sigma^{-1}\right),$$

$$D_{\Sigma A} \log p\left(\boldsymbol{y}|\Sigma, A\right) = D_{A\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right),$$

$$D_{\Sigma\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) = K_{NN}\left\{\begin{array}{c} \frac{T}{2}\frac{1}{2}\left(\left(\Sigma^{-1}\right)' \otimes \Sigma^{-1} + \left(\Sigma^{-1}\right)' \otimes \Sigma^{-1}\right) \\ -\frac{1}{2}\left(\begin{array}{c} \left(\Sigma^{-1}\left(\boldsymbol{y} - XA\right)'\left(\boldsymbol{y} - XA\right)\Sigma^{-1}\right)' \otimes \Sigma^{-1} \\ +\left(\Sigma^{-1}\right)' \otimes \left(\Sigma^{-1}\left(\boldsymbol{y} - XA\right)'\left(\boldsymbol{y} - XA\right)\Sigma^{-1}\right) \end{array}\right) \end{array}\right\},$$

where where K_{NN} is the commutation matrix for a matrix with N rows and N columns. Since Σ is symmetric, define $V\Sigma = vech\left(\Sigma\right)$ and we have an index matrix $D_{V\Sigma}\left(\Sigma\right)$ which is defined as

$$D_{V\Sigma}\left(\Sigma\right) = diag\left(R_1, R_2, ...R_m, ..., R_N\right).$$

where

$$R_k = \left[\begin{array}{c} 0_{(m-1)\times(N-m+1)} \\ I_{N-m+1} \end{array}\right]_{N\times(N-m+1)}.$$

Hence, the derivatives are

$$\begin{aligned} D_A \log p\left(\boldsymbol{y}|\Sigma, A\right) &= \left[vec\left(\left(\frac{1}{2}\left(\Sigma^{-1\prime} + \Sigma^{-1}\right)\left(\boldsymbol{y} - XA\right)' X\right)'\right)\right]' \\ D_{V\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) &= D_\Sigma \log p\left(\boldsymbol{y}|\Sigma, A\right) \times D_{V\Sigma}\left(\Sigma\right) \end{aligned}$$

$$\begin{aligned} D_{AA} \log p\left(\boldsymbol{y}|\Sigma, A\right) &= -\frac{1}{2}\left(\left(\Sigma^{-1\prime} + \Sigma^{-1}\right) \otimes X'X\right), \\ D_{AV\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) &= D_{A\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) \times D_{V\Sigma}\left(\Sigma\right), \\ D_{V\Sigma A} \log p\left(\boldsymbol{y}|\Sigma, A\right) &= D'_{AV\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right), \\ D_{V\Sigma V\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) &= \left(D_{V\Sigma}\left(\Sigma\right)' \otimes I_1\right) \times D_{\Sigma\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) \times D_{V\Sigma}\left(\Sigma\right) \\ &= D_{V\Sigma}\left(\Sigma\right)' \times D_{\Sigma\Sigma} \log p\left(\boldsymbol{y}|\Sigma, A\right) \times D_{V\Sigma}\left(\Sigma\right). \end{aligned}$$

6.2 Appendix 2: The derivation of RDIC for the factor VAR model

The model can be written in matrix form

$$\begin{aligned} y_t &= L\boldsymbol{f}_t + e_t, \\ \boldsymbol{f}_t &= \Phi F_{t-1} + \varepsilon_t, \end{aligned}$$

where $\Phi = (\Phi_1, \Phi_2, ..., \Phi_h)$, $\boldsymbol{F}_t = \left(\boldsymbol{f}'_t, ..., \boldsymbol{f}'_{t-h+1}\right)'$.

The complete data log-likelihood function is:

$$\begin{aligned} \log f\left(\boldsymbol{y}, \boldsymbol{f} \,|L, \Sigma, \Phi, Q\right) &= -\frac{NT + q\left(T-h\right)}{2}\log 2\pi - \frac{T}{2}\log|\Sigma| \\ &\quad -\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\right] \\ &\quad -\frac{T-h}{2}\log|Q| - \frac{1}{2}\mathbf{tr}\left[Q^{-1}\left(\boldsymbol{f}_{+\mathbf{1}} - \boldsymbol{f}_{-1}\Phi'\right)'\left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1}\Phi'\right)\right], \end{aligned}$$

where $\boldsymbol{f}_{+1} = \left[\boldsymbol{f}_{h+1}, \boldsymbol{f}_{h+2}, ..., \boldsymbol{f}_T\right]'$, $\boldsymbol{f}_{-1} = [\boldsymbol{F}_h, \boldsymbol{F}_{h+1}, ..., \boldsymbol{F}_{T-1}]'$. Denote this function by $\varphi(L, \Sigma, \Phi, Q)$. We now derive the first and second derivative of it.

The first order derivatives of $\varphi(L, \Sigma, \Phi, Q)$:

Whenever there is no confusion, we denote $\varphi(L, \Sigma, \Phi, Q)$ simply by φ. The derivative of $\varphi(L, \Sigma, \Phi, Q)$ with respect to L is

$$\begin{aligned} d_L(\varphi) &= d\left(-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\right]\right) \\ &= -\frac{1}{2}\mathbf{tr}\left\{-\Sigma^{-1}\left(dL\right)\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right) + \Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)'\left(-\boldsymbol{f}\left(dL\right)'\right)\right\} \\ &= \frac{1}{2}\mathbf{tr}\left\{\Sigma^{-1}dL\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right) + \Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)'\boldsymbol{f}\left(dL\right)'\right\} \\ &= \frac{1}{2}\mathbf{tr}\left\{\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\Sigma^{-1}dL + dL\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\Sigma^{-1}\right)'\right\} \\ &= \frac{1}{2}\mathbf{tr}\left\{\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\left(\Sigma^{-1}\right)' + \Sigma^{-1}\right)dL\right\} \\ &= \mathbf{tr}\left(\widetilde{c}dL\right), \end{aligned}$$

where

$$\widetilde{c} = \frac{1}{2}\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\left(\Sigma^{-1}\right)' + \Sigma^{-1}\right).$$

Taking the *vec* operation, we get

$$d\left(vec\left(-\frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}(\boldsymbol{y} - \boldsymbol{f}L')'(\boldsymbol{y} - \boldsymbol{f}L')\right]\right)\right) = d(vec(\varphi)) = \left(vec(\widetilde{c})'\right)' d(vec(L)).$$

The first derivative of $\varphi\left(L, \Sigma, \Phi, Q\right)$ is

$$D_L\left(\varphi\right) = \left(vec\left(\left[\frac{1}{2}\boldsymbol{f}'\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\left(\Sigma^{-1}\right)' + \Sigma^{-1}\right)\right]'\right)\right)'.$$

Similarly, we have

$$D_{\Sigma}\left(\varphi\right) = \left(vec\left(-\frac{T-s}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-1}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)^{'}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\Sigma^{-1}\right)^{'}\right)^{'},$$

$$D_{\Phi}\left(\varphi\right) = \left(vec\left(\left[\frac{1}{2}\boldsymbol{f}'_{-1}\left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1}\Phi'\right)\left(\left(Q^{-1}\right)^{'} + Q^{-1}\right)\right]^{'}\right)\right)^{'},$$

$$D_{Q}\left(\varphi\right) = \left(vec\left(-\frac{T-h}{2}Q^{-1} + \frac{1}{2}Q^{-1}\left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1}\Phi'\right)^{'}\left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1}\Phi'\right)Q^{-1}\right)^{'}\right)^{'}.$$

The second order derivatives of $\varphi\left(L, \Sigma, \Phi, Q\right)$:

The first order derivative of $\widetilde{c}$ is

$$d\widetilde{c} = d\left(\frac{1}{2}\boldsymbol{f}^{'}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\left(\Sigma^{-1}\right)^{'} + \Sigma^{-1}\right)\right) = -\frac{1}{2}\boldsymbol{f}^{'}\boldsymbol{f}\left(dL\right)^{'}\left(\left(\Sigma^{-1}\right)^{'} + \Sigma^{-1}\right).$$

And the second order derivative is

$$\begin{aligned} d_{L}^{2}\varphi &= \mathbf{tr}\left(d\widetilde{c} * dL\right) \\ &= \mathbf{tr}\left(-\frac{1}{2}\boldsymbol{f}^{'}\boldsymbol{f}\left(dL\right)^{'}\left(\left(\Sigma^{-1}\right)^{'} + \Sigma^{-1}\right)dL\right). \end{aligned}$$

Then, we have,

$$D_{L,L}\left(\varphi\right) = -\frac{1}{2}\left(\boldsymbol{f}^{'}\boldsymbol{f} \otimes \left(\left(\Sigma^{-1}\right)^{'} + \Sigma^{-1}\right)\right),$$

$$\begin{aligned} H &= G\left(T\right) = T^{'}, \quad T = S\left(\Sigma\right) = \frac{1}{2}\boldsymbol{f}^{'}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\left(\left(\Sigma^{-1}\right)^{'} + \Sigma^{-1}\right), \\ D\left(G\left(T\right)\right) &= K_{qN}, \\ D\left(S\left(\Sigma\right)\right) &= I_{N} \otimes \left(\boldsymbol{f}^{'}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\right) \cdot \left(-\frac{1}{2}\left(K_{NN} + I_{NN}\right)\right) \cdot \left(\left(\Sigma^{-1}\right)^{'} \otimes \Sigma^{-1}\right), \\ DH\left(\Sigma\right) &= \left(DG\left(T\right)\right)\left(DS\left(\Sigma\right)\right), \end{aligned}$$

where K_{qM} is the commutation matrix for a matrix with q rows and N columns. Thus, we have

$$\begin{aligned} D_{L,\Sigma}\left(\varphi\right) &= \frac{\partial D_{L}\left(\varphi\right)}{\left(\partial vec\Sigma\right)^{'}} = \left(DG\left(T\right)\right)\left(DS\left(\Sigma\right)\right) \\ &= K_{qN} \cdot I_{N} \otimes \left(\boldsymbol{f}^{'}\left(\boldsymbol{y} - \boldsymbol{f}L'\right)\right) \cdot \left(-\frac{1}{2}\left(K_{NN} + I_{NN}\right)\right) \cdot \left(\left(\Sigma^{-1}\right)^{'} \otimes \Sigma^{-1}\right), \end{aligned}$$

$$\begin{aligned} D_{L,\Phi}\left(\varphi\right) &= 0, \\ D_{L,Q}\left(\varphi\right) &= 0, \end{aligned}$$

$$D_{\Sigma,\Sigma}(\varphi) = K_{NN} \cdot \left(\begin{array}{c} \frac{T-s}{2} \cdot \frac{1}{2} \left(\left(\Sigma^{-1}\right)' \otimes \Sigma^{-1} + \left(\Sigma^{-1}\right)' \otimes \Sigma^{-1} \right) \\ -\frac{1}{2} \left(\begin{array}{c} \left(\Sigma^{-1} \left(\boldsymbol{y} - \boldsymbol{f} L'\right)' \left(\boldsymbol{y} - \boldsymbol{f} L'\right) \Sigma^{-1} \right)' \otimes \Sigma^{-1} \\ + \left(\Sigma^{-1}\right)' \otimes \left(\Sigma^{-1} \left(\boldsymbol{y} - \boldsymbol{f} L'\right)' \left(\boldsymbol{y} - \boldsymbol{f} L'\right) \Sigma^{-1} \right) \end{array} \right) \end{array} \right),$$

$$\begin{aligned} D_{\Sigma,\Phi}(\varphi) &= 0, \\ D_{\Sigma,Q}(\varphi) &= 0, \end{aligned}$$

$$\begin{aligned} & D_{\Phi,Q}(\varphi) \\ = \;& K_{qq} \cdot \left(I_K \otimes \boldsymbol{f}'_{-1} \left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1} \Phi' \right) \right) \cdot \left(-\frac{1}{2} \left(K_{qq} + I_{qq} \right) \right) \cdot \left(\left(Q^{-1}\right)' \otimes Q^{-1} \right), \end{aligned}$$

$$D_{\Phi,\Phi}(\varphi) = -\frac{1}{2} \left(\boldsymbol{f}'_{-1} \boldsymbol{f}_{-1} \otimes \left(\left(Q^{-1}\right)' + Q^{-1} \right) \right),$$

$$D_{Q,Q}(\varphi) = K_{KK} \left(\frac{T-s}{4} \left(\left(Q^{-1}\right)' \otimes Q^{-1} + \left(Q^{-1}\right)' \otimes Q^{-1} \right) - \frac{1}{2} M \right),$$

where

$$\begin{aligned} M = \;& Q^{-1} \left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1} \Phi' \right)' \left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1} \Phi' \right) Q^{-1'} \otimes Q^{-1} + \\ & \left(Q^{-1}\right)' \otimes \left(\Sigma^{-1} \left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1} \Phi' \right)' \left(\boldsymbol{f}_{+1} - \boldsymbol{f}_{-1} \Phi' \right) Q^{-1} \right). \end{aligned}$$

The special structure of the parameter matrices:

Let

$$L^* = vec\left(\overline{L}\right), \Sigma^* = diag\left(\Sigma\right),\ \Phi^* = vec\left(\Phi\right),\ Q^* = vech\left(Q\right),$$

where $\overline{L}$ is the last $(N-q) \times q$ block of L. We now obtain the derivative of these parameter matrices.

The first order derivatives are as follows:

$$\begin{aligned} D_{L^*}(\varphi) &= D_L(\varphi) \cdot D_{L^*}\left(L\left(L^*\right)\right) = D_L(\varphi) \cdot \dot{I}_{L^*}, \\ D_{\Sigma^*}(\varphi) &= D_\Sigma(\varphi) \cdot D_{\Sigma^*}\left(\Sigma\left(\Sigma^*\right)\right) = D_\Sigma(\varphi) \cdot \dot{I}_{\Sigma^*}, \\ D_{\Phi^*}(\varphi) &= D_\Phi(\varphi) \cdot \dot{I}_{\Phi^*}, \\ D_{Q^*}(\varphi) &= D_Q(\varphi) \cdot \dot{I}_{Q^*}. \end{aligned}$$

The second order derivatives are as follows:

$$\begin{aligned}
D_{L^*,L^*}(\varphi) &= D_{L^*}(D_{L^*}(\varphi)) = D_{L^*}\left(D_L(\varphi)\cdot \dot{I}_{L^*}\right) \\
&= \left(\dot{I}_{L^*}' \otimes I_1\right)\cdot D_{L^*}(D_L(\varphi)) \\
&= \left(\dot{I}_{L^*}' \otimes I_1\right)\cdot D_{L,L}(\varphi)\cdot \dot{I}_{L^*}, \\
D_{L^*,\Sigma^*}(\varphi) &= D_{\Sigma^*}(D_{L^*}(\varphi)) = D_{\Sigma^*}\left(D_L(\varphi)\cdot \dot{I}_{L^*}\right) \\
&= \left(\dot{I}_{L^*}' \otimes I_1\right)\cdot D_{\Sigma^*}(D_L(\varphi)) \\
&= \left(\dot{I}_{L^*}' \otimes I_1\right)\cdot D_{\Sigma}(D_L(\varphi))\cdot D_{\Sigma^*}(\Sigma(\Sigma^*)) \\
&= \dot{I}_{L^*}'\cdot D_{L,\Sigma}(\varphi)\cdot \dot{I}_{\Sigma^*}, \\
D_{L^*,\Phi^*}(\varphi) &= 0, \\
D_{L^*,Q^*}(\varphi) &= 0,
\end{aligned}$$

$$\begin{aligned}
D_{\Sigma^*,\Sigma^*}(\varphi) &= D_{\Sigma^*}(D_{\Sigma^*}(\varphi)) = D_{\Sigma^*}\left(D_\Sigma(\varphi)\cdot \dot{I}_{\Sigma^*}\right) \\
&= \dot{I}_{\Sigma^*}' \otimes I_1\cdot D_{\Sigma^*}(D_\Sigma(\varphi)) \\
&= \dot{I}_{\Sigma^*}'\cdot D_\Sigma(D_\Sigma(\varphi))\cdot \dot{I}_{\Sigma^*}, \\
D_{\Sigma^*,\Phi^*}(\varphi) &= 0, \\
D_{\Sigma^*,Q^*}(\varphi) &= 0.
\end{aligned}$$

$$\begin{aligned}
D_{\Phi^*,\Phi^*}(\varphi) &= \dot{I}_{\Phi^*}'\cdot (D_{\Phi,\Phi}(\varphi))\cdot \dot{I}_{\Phi^*}, \\
D_{\Phi^*,Q^*}(\varphi) &= \dot{I}_{\Phi^*}'\cdot (D_{\Phi,Q}(\varphi))\cdot \dot{I}_{Q^*}, \\
D_{Q^*,Q^*}(\varphi) &= \dot{I}_{Q^*}'\cdot D_{Q,Q}(\varphi)\cdot \dot{I}_{Q^*},
\end{aligned}$$

where $D_{L^*}(L(L^*)) = \dot{I}_{L^*}$, $D_{\Sigma^*}(\Sigma(\Sigma^*)) = \dot{I}_{\Sigma^*}$.

For $\dot{I}_{L^*}$ which is a block diagonal matrix, we have

$$\dot{I}_{L^*} = diag(P_1, P_2, ..., P_q),$$

where

$$P_i = \begin{bmatrix} 0_{q\times(N-q)} \\ I_{N-q} \end{bmatrix}.$$

And for $\dot{I}_{\Sigma^*}$, which is an $N^2 \times N$ matrix whose n^{th} column has 1 in the $((n-1)\times N + n)^{th}$ row and other elements are all zeros. For $\dot{I}_{\Phi^*}$, we have

$$\dot{I}_{\Phi^*} = I_{q*q}.$$

For $\dot{I}_{Q^*}$, we have

$$\dot{I}_{Q^*} = diag\left(R_1, R_2, ...R_k, ..., R_q\right).$$

where

$$R_k = \begin{bmatrix} 0_{(k-1)\times(q-k+1)} \\ I_{q-k+1} \end{bmatrix}_{q\times(q-k+1)},$$

since Q is symmetric.

The first order derivative matrix of the complete-data likelihood with respect to $L^*, \Sigma^*, \Phi^*, Q^*$ is:

$$vec\left(\begin{bmatrix} D_{L^*}(\varphi) & D_{\Sigma^*}(\varphi) & D_{\Phi^*}(\varphi) & D_{Q^*}(\varphi) \end{bmatrix}\right).$$

The second order derivative matrix of the complete data likelihood with respect to $L^*, \Sigma^*, \Phi^*, Q^*$ is:

$$\begin{bmatrix} D_{L^*,L^*}(\varphi) & D_{L^*,\Sigma^*}(\varphi) & 0 & 0 \\ D_{\Sigma^*,L^*}(\varphi) & D_{\Sigma^*,\Sigma^*}(\varphi) & 0 & 0 \\ 0 & 0 & D_{\Phi^*,\Phi^*}(\varphi) & D_{\Phi^*,Q^*}(\varphi) \\ 0 & 0 & D_{Q^*,\Phi^*}(\varphi) & D_{Q^*,Q^*}(\varphi) \end{bmatrix}.$$

References

[1] Akaike, H. (1973): "Information theory and an extension of the maximum likelihood principle," in *Second international symposium on information theory*, Springer Verlag, Vol 1, 267–281.

[2] Bai, J. and P. Wang (2012): "Identification and estimation of dynamic factor models," *Working Paper*.

[3] Bernanke, B. S. and J. Boivin (2003): "Monetary policy in a data-rich environment," *Journal of Monetary Economics*, 50, 525–546.

[4] Bernanke, B. S., J. Boivin, and P. Eliasz (2005): "Measuring the effects of monetary policy: A factor-augmented vector autoregressive (FAVAR) approach," *The Quarterly Journal of Economics*, 120, 387–422.

[5] Boivin, J. and S. Ng (2005): "Understanding and comparing factor-based forecasts," *International Journal of Central Banking*, 1, 117–151

[6] Chan, J. C. C. and A. L. Grant (2014): "Issues in Comparing Stochastic Volatility Models Using the Deviance Information Criterion," *Working Paper*, 2014.

[7] Chen, C. F. (1985): "On Asymptotic Normality of Limiting Density Functions with Bayesian Implications," *Journal of the Royal Statistical Society, Series B*, 47, 540–546.

[8] Forni, M., M. Hallin, M. Lippi, and L. Reichlin (2003): "Do financial variables help forecasting inflation and real activity in the euro area?" *Journal of Monetary Economics*, 50, 1243–1255.

[9] Kass, R. E. and A. E. Raftery (1995): "Bayes factors," *Journal of the American Statistical Association*, 90, 773–795.

[10] Koop, G. and D. Korobilis (2009): "Bayesian multivariate time series methods for empirical macroeconomics," *Foundations and Trends® in Econometrics*, 3, 267–358.

[11] Koop, G. and S. M. Potter (2011): "Time varying VARs with inequality restrictions," *Journal of Economic Dynamics and Control*, 35, 1126–1138.

[12] Kose, M. A., C. Otrok, and C. H. Whiteman (2003): "International business cycles: World, region, and country-Specific Factors," *American Economic Review*, 93, 1216–1239.

[13] Kose, M. A., C. Otrok, and C. H. Whiteman (2008): "Understanding the evolution of world business cycles," *Journal of International Economics*, 75, 110–130.

[14] Li, Y. and J. Yu (2012): "Bayesian hypothesis testing in latent variable models," *Journal of Econometrics*, 166, 237–246.

[15] Li, Y., T. Zeng, and J. Yu (2014a): "A new approach to Bayesian hypothesis testing," *Journal of Econometrics*, 178, 602–612.

[16] Li, Y., T. Zeng, and J. Yu (2014b): "Robust deviation information criterion for latent variable models," *Working Paper*, Singapore Management University.

[17] Louis, T. A. (1982): "Finding the observed information matrix when using the EM Algorithm," *Journal of the Royal Statistical Society, Series B*, 44, 226–233.

[18] Magnus, J. and H. Neudecker (1999): *Matrix Differential Calculus with Applications in Statistics and Econometrics*, Wiley, Chichester.

[19] Ni, S. and D. Sun (2003): "Noninformative priors and frequentist risks of Bayesian estimators of vector-autoregressive models," *Journal of Econometrics*, 115, 159–197.

[20] Otrok, C. and C. H. Whiteman (1998): "Bayesian leading indicators: measuring and predicting economic conditions in Iowa," *International Economic Review*, 997–1014.

[21] Phillips, P.C.B. (1995): "Bayesian model selection and prediction with empirical applications," *Journal of Econometrics*, 69, 289–331.

[22] Phillips, P.C.B. (1996): "Econometric Model Determination," *Econometrica*, 64, 763–812.

[23] Robert, C. (2001): *The Bayesian Choice: From Decision-Theoretic Foundations to Computational Implementation*, Springer.

[24] Schwarz, G. (1978): "Estimating the dimension of a model," *Annals of Statistics*, 6(2), 461-464.

[25] Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002): "Bayesian measures of model complexity and fit," *Journal of the Royal Statistical Society: Series B*, 64, 583–639.

[26] Stock, J. H. and M. W. Watson (2002a): "Forecasting using principal components from a large number of predictors," *Journal of the American Statistical Association*, 97, 1167–1179.

[27] Stock, J. H. and M. W. Watson (2002b): "Macroeconomic forecasting using diffusion indexes," *Journal of Business & Economic Statistics*, 20, 147–162.

[28] Tanner, M. and W. Wong (1987): "The calculation of posterior distributions by data augmentation," *Journal of the American Statistical Association*, 82, 528–540.