# Robust Deviance Information Criterion for Latent Variable Models*

Yong Li
*Renmin University of China*

Tao Zeng
*Wuhan University*

Jun Yu
*Singapore Management University*

December 10, 2015

**Abstract**

Deviance information criterion (DIC) is a widely used information criterion for Bayesian model comparison. In this paper a rigorous decision-theoretic justification of DIC is provided for models without latent variables or incidental parameters. For models with latent variables, however, it is shown that the data augmentation technique undermines the theoretical underpinnings of DIC, although it facilitates parameter estimation via Markov chain Monte Carlo (MCMC) simulation. Data augmentation invalidates the standard asymptotic arguments and conventional estimators of latent variables may be inconsistent. In this paper, a robust form of DIC, denoted as RDIC, is advocated for Bayesian comparison of latent variable models. RDIC is shown to be a good approximation to DIC without data augmentation. While the later quantity is difficult to compute, the expectation – maximization (EM) algorithm facilitates the computation of RDIC when the MCMC output is available. Moreover, RDIC is robust to nonlinear transformations of latent variables and distributional representations of model specification. The proposed approach is applied to several popular models in economics and finance.

*JEL classification:* C11, C12, G12
*Keywords:* AIC; DIC; EM Algorithm; Latent variable models; Markov Chain Monte Carlo.

## 1 Introduction

One of the most important developments in the Bayesian literature in recent years is arguably the deviance information criterion (DIC) of Spiegelhalter, et al (2002). DIC is a Bayesian

*We wish to thank Eric Renault (co-editor), two referees, Peter Phillips and David Spiegelhalter for their helpful comments. Yong Li, Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing, 100872, P.R. China. Tao Zeng, Economics and Management School, Wuhan University, Wuhan, China 430072. Jun Yu, School of Economics and Lee Kong Chian School of Business, Singapore Management University, 90 Stamford Rd, Singapore 178903. Email for Jun Yu: yujun@smu.edu.sg. URL: http://www.mysmu.edu/faculty/yujun/. Yu thanks the Singapore Ministry of Education for Academic Research Fund under grant number MOE2011-T2-2-096.

version of the well known Akaike Information Criterion (AIC) (Akaike, 1973). Like AIC, it trades off a measure of model adequacy against a measure of complexity and is concerned with how replicate data predict the observed data. DIC is constructed based on the posterior distribution of the log-likelihood or the deviance, and has several desirable features. Firstly, DIC is simple to calculate when the likelihood function is available in closed-form and the posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. Secondly, it is applicable to a wide range of statistical models. Third, unlike Bayes factors (BFs), it is not subject to Jeffery-Lindley's paradox. However, as acknowledged by Spiegelhalter, et al (2002, 2014), so far there is no rigorous decision-theoretic justification of DIC in the literature. The first contribution of the present paper is to provide a rigorous justification when the standard Bayesian large sample theory is valid.

An important class of models in economics and finance involves latent variables. Latent variables have figured prominently in consumption decision, investment decision, labor force participation, conduct of monetary policy, indices of economic activity, inflation dynamics and other economic, business and financial activities and decisions. Not surprisingly, latent variable models have been widely used in financial econometrics, macroeconometrics and microeconometrics. For example, in financial econometrics it is often found that the values of stocks, bonds, options, futures, and derivatives are often determined by a small number of factors. These factors, such as the level, the slope and the curvature in the term structure of interest rates, are latent. In macroeconomics, a well-known recent example of latent variable models is the dynamic stochastic general equilibrium (DSGE) model. On the basis of macroeconomic theory, the DSGE model attempts to explain aggregate economic phenomena by taking into account the fact that the economy is affected by some structural innovations. The DSGE model can be solved as a rational expectation system in the percentage deviation of variables from their steady-states which are latent, see An and Schorfheide (2007) and Dejong and Dave (2007). In microeconometrics, many discrete choice models and panel data models involve unobserved variables in order to capture observed heterogeneity across economic entities (Stern, 1997).

For latent variable models, Bayesian methods via MCMC simulation have proven to be a powerful alternative to frequentist methods for estimating model parameters. In particular, the *data augmentation* strategy proposed by Tanner and Wong (1987), that expands the parameter space by treating the latent variables as additional model parameters, has been found very useful for simplifying the MCMC computation of posterior distributions. This simplification is achieved because data augmentation leads to a closed-form expression for the likelihood function.

Comparing alternative latent variable models in the Bayesian paradigm is a daunting and yet important task. The gold standard to carry out Bayesian model comparison is to compute BFs, which basically compare marginal likelihood of alternative models (Kass and

Raftery, 1995). Several interesting developments have been made in recent years for computing marginal likelihood from the MCMC output; see for example, Chib (1995), Chib and Jeliazkov (2001). While these methods are very general and widely applicable, for latent variable models, they are difficult to use because the marginal likelihood may be hard to calculate. In addition, BFs cannot be used under improper priors and are subject to the Jeffrey-Lindley paradox. Given that DIC is simple to calculate from the MCMC output with the data augmentation technique and also that data augmentation is often used for Bayesian parameter estimation, DIC has been widely used for comparing alternative latent variable models; see for example, Berg, Meyer and Yu (2004), Huang and Yu (2010).

The second contribution of the present paper is that we argue that DIC has to be used with care in the context of latent variable models. In particular, we believe DIC, in the way in which it is commonly implemented, has some conceptual and practical problems. Firstly, when the latent variables are treated as parameters, the standard Bayesian large sample theory is not applicable and hence DIC is not asymptotically justified. Secondly, DIC is not robust to apparently innocuous transformations and distributional representations. This problem is made worse by the data augmentation technique for latent variable models. Data augmentation greatly inflates the number of parameters and hence the "effective" number of parameters used in DIC is very sensitive to transformations and distributional representations. Without data augmentation, however, the likelihood function that is based on observed data only, does not have a closed-form expression and hence the corresponding DIC is much harder to compute for latent variable models.

The third contribution of the present paper is that we advocate the use of a robust version of DIC, denoted by RDIC, to make Bayesian comparison of latent variable models. It is shown that RDIC is a good approximation to DIC without data augmentation and both are asymptotically justified. We then show that the expectation – maximization (EM) algorithm facilitates the computation of RDIC for latent variable models when the MCMC output is available. Moreover, RDIC is robust to nonlinear transformations of latent variables and to distributional representations of model specification.

The advantages of the proposed approach are illustrated using two popular models in economics and finance, including a class of dynamic factor models and a class of stochastic volatility models. It is shown that if the latent variables are treated as parameters, DIC is very sensitive to the nonlinear transformations of latent variables in these models, whereas RDIC is robust to these transformations. As a result, substantial discrepancy is found between DIC and RDIC.

The paper is organized as follows. In Section 2, the latent variable models are introduced. The Bayesian estimation method with data augmentation and the EM algorithm are also reviewed. Section 3 provides a rigorous decision-theoretic justification of DIC for models without latent variable under a set of regularity conditions. In Section 4, we show that the

commonly used version of DIC is not justified for models with latent variable models. We also introduce RDIC to compare latent variable model and discuss how to compute RDIC from the MCMC output. Section 5 illustrates the method using models from economics and finance. Section 6 concludes the paper. The Appendix collects the proof of the theoretical results in the paper.

## 2 Latent Variable Models, EM Algorithm and MCMC

Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ denote observed variables and $\mathbf{z} = (z_1, z_2, \cdots, z_n)'$ the latent variables. The latent variable model is indexed by the a set of $P$ parameters, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)' \in \boldsymbol{\Theta} \subseteq R^P$. Let $p(\mathbf{y}|\boldsymbol{\theta})$ be the likelihood function of the observed data (denoted the observed-data likelihood), and $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ be the complete-data likelihood function. The relationship between the two functions is:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}. \tag{1}$$

In many cases, the integral does not have a closed-form solution. Consequently, statistical inferences, such as estimation and model comparison, are difficult to make. In the literature, maximum likelihood (ML) analysis using the EM algorithm and Bayesian analysis using MCMC are two popular approaches for carrying out statistical inference of the latent variable models.

### 2.1 Maximum likelihood via the EM algorithm

The EM algorithm is an iterative numerical method for finding the ML estimates of $\boldsymbol{\theta}$ in the latent variable models. It has been widely used in applications since Dempster, et al (1977) gave its name and did the convergence analysis. In this subsection, we briefly review the main idea of the EM algorithm. For more details, one can refer to McLachlan and Krishnan (2008).

Let $\mathbf{x} = (\mathbf{y}, \mathbf{z})$ be the complete data with a density $p(\mathbf{x}|\boldsymbol{\theta})$. The observed-data log-likelihood $\mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta}) = \ln p(\mathbf{y}|\boldsymbol{\theta})$ often involves some intractable integral, preventing researchers from directly optimizing $\mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. In many cases, however, the complete-data log-likelihood $\mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta}) = \ln p(\mathbf{x}|\boldsymbol{\theta})$ has a closed-form expression. Instead of maximizing $\mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta})$ directly, the EM algorithm maximizes $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$, the conditional expectation of the complete-data log-likelihood function $\mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})$ given the observed data $\mathbf{y}$ and a current fit $\boldsymbol{\theta}^{(r)}$ of the parameter.

Generally, a standard EM algorithm has two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = E_{\mathbf{z}}\{\mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})|\mathbf{y}, \theta^{(r)}\}, \tag{2}$$

where the expectation is taken with respect to the conditional distribution $p(\mathbf{z}|\mathbf{y}, \boldsymbol{\theta}^{(r)})$. The M-step determines a $\boldsymbol{\theta}^{(r+1)}$ that maximizes $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$. Under some mild regularity conditions,

the sequence $\{\boldsymbol{\theta}^{(r)}\}$ obtained from the EM iterations converges to the ML estimate $\hat{\boldsymbol{\theta}}$; see Dempster, et al (1977) and Wu (1983) for details on the convergence properties of $\{\boldsymbol{\theta}^{(r)}\}$.

## 2.2 Bayesian analysis using MCMC

Although the EM algorithm is a reasonable statistical approach for analyzing latent variable models, the numerical optimization in the $M$-step is often unstable. This numerical problem worsens as the dimension of $\boldsymbol{\theta}$ increases. It is well recognized that Bayesian methods using MCMC provide a powerful tool to analyze the latent variables models. However, if the posterior analysis is conducted from the observed-data likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, one would end up with the same problem as in the ML method since $p(\mathbf{y}|\boldsymbol{\theta})$ does not have a closed-form expression.

The novelty in the Bayesian methods is to treat the latent variable model as a hierarchical structure of conditional distributions, namely, $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$, $p(\mathbf{z}|\boldsymbol{\theta})$, and $p(\boldsymbol{\theta})$. In other words, one can use the data augmentation strategy of Tanner and Wong (1987) to expand the parameter space from $\boldsymbol{\theta}$ to $(\boldsymbol{\theta},\mathbf{z})$. The advantage of data augmentation is that the Bayesian analysis is now based on the new likelihood function, $p(\mathbf{y}|\boldsymbol{\theta},\mathbf{z})$ which often has a closed-form expression. Then the Gibbs sampler and other MCMC samplers can be used to generate random samples from the joint posterior distribution $p(\boldsymbol{\theta},\mathbf{z}|\mathbf{y})$. After a sufficiently long period for a burning-in phase, the simulated random samples can be regarded as random observations from the joint distribution. The statistical analysis can be established on the basis of these simulated posterior random observations. As a by-product to the Bayesian analysis, one also obtains Markov chains for the latent variables $\mathbf{z}$ and hence statistical inference can be made about $\mathbf{z}$. For further details on Bayesian analysis of latent variable models via MCMC, including algorithms, examples and references, see Geweke, et al (2011). From the above discussion, it can be seen that data augmentation is the key technique for Bayesian estimation of latent variable models.

Two observations are in order. First, with data augmentation, the parameter space is much bigger. More often than not, the dimension of the space increases as the number of observations increases and is larger than the number of observations. In this case, the new likelihood function is not regular. Second, it is difficult to argue that the latent variables can be always treated as model parameters. Models parameters are typically fixed but the latent variables are often time varying. Consequently, the same treatment of these two types of variables does not seem to be justifiable from the perspective of model selection.

# 3 Decision-theoretic Justification of DIC

## 3.1 DIC

We first review DIC for models without latent variables. Spiegelhalter, et al (2002) proposed DIC for Bayesian model comparison. The criterion is based on the deviance

$$D(\boldsymbol{\theta}) = -2\ln p(\mathbf{y}|\boldsymbol{\theta}),$$

and takes the form of[1]

$$\mathrm{DIC}_1 = \overline{D(\boldsymbol{\theta})} + P_D. \tag{3}$$

The first term, used as a Bayesian measure of model fit, is defined as the posterior expectation of the deviance, that is,

$$\overline{D(\boldsymbol{\theta})} = E_{\boldsymbol{\theta}|\mathbf{y}}[D(\boldsymbol{\theta})] = E_{\boldsymbol{\theta}|\mathbf{y}}[-2\ln p(\mathbf{y}|\boldsymbol{\theta})].$$

The better the model fits the data, the larger the log-likelihood value and hence the smaller the value for $\overline{D(\boldsymbol{\theta})}$. The second term, used to measure the model complexity and also known as "effective number of parameters", is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters:

$$P_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) = -2\int [\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})]p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}, \tag{4}$$

where $\bar{\boldsymbol{\theta}}$ is the Bayesian estimator, and more precisely the posterior mean, of the parameter $\boldsymbol{\theta}$. Here, $P_D$ can be explained as the expected excess of the true over the estimated residual information conditional on data $\mathbf{y}$. In other words, $P_D$ can be interpreted as the expected reduction in uncertainty due to estimation.

DIC can be rewritten by two equivalent forms:

$$\mathrm{DIC}_1 = D(\bar{\boldsymbol{\theta}}) + 2P_D, \tag{5}$$

and

$$\mathrm{DIC}_1 = 2\overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) = -4E_{\boldsymbol{\theta}|\mathbf{y}}[\ln p(\mathbf{y}|\boldsymbol{\theta})] + 2\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}). \tag{6}$$

$\mathrm{DIC}_1$ defined in Equation (5) bears similarity to AIC of Akaike (1973) and can be interpreted as a classical "plug-in" measure of fit plus a measure of complexity (i.e., $2P_D$, also known as the penalty term). In Equation (3) the Bayesian measure, $\overline{D(\boldsymbol{\theta})}$, is the same as $D(\bar{\boldsymbol{\theta}}) + P_D$ which already includes a penalty term for model complexity and thus could be better thought of as a measure of model adequacy rather than pure goodness of fit.

To calculate DIC, one needs to determine the likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, and hence the parameters, $\boldsymbol{\theta}$. In some cases, a clear definition of parameters may not be obvious. Taking a hierarchical

[1] We use $\mathrm{DIC}_1$ instead of DIC because for latent variable models other forms of DIC are possible and will be discussed later.

model as an example. Let $p(\mathbf{y}|\boldsymbol{\theta})$ be a hierarchical model with a prior distribution $p(\boldsymbol{\theta}|\boldsymbol{\psi})$ where $\boldsymbol{\psi}$ is assigned with another prior distribution. In this case, there are two ways to define parameters and the likelihood. The idea may be explained using an important concept, namely, focus. Spiegelhalter et al. (2002, pages 584-585) introduced focus to help determine the parameters, the likelihood function and DIC. They showed that if $\boldsymbol{\theta}$ is identified as parameters in focus in the hierarchical model, we should use $p(\mathbf{y}|\boldsymbol{\theta})$ to construct DIC, i.e.,

$$D(\bar{\boldsymbol{\theta}}) = -2\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}),\ P_D = -2\int \left[\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})\right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta},$$
$$\mathrm{DIC}_1 = D(\bar{\boldsymbol{\theta}}) + 2P_D.$$

The prior distribution $p(\boldsymbol{\theta})$ can be obtained from,

$$p(\boldsymbol{\theta}) = \int p(\boldsymbol{\theta}|\boldsymbol{\psi})p(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi}.$$

If $\boldsymbol{\psi}$ is identified as parameters in focus, we should then use $p(\mathbf{y}|\boldsymbol{\psi})$ to construct DIC, i.e.,

$$D(\bar{\boldsymbol{\theta}}) = -2\ln p(\mathbf{y}|\bar{\boldsymbol{\psi}}), P_D = -2\int \left[\ln p(\mathbf{y}|\boldsymbol{\psi}) - \ln p(\mathbf{y}|\bar{\boldsymbol{\psi}})\right] p(\boldsymbol{\psi}|\mathbf{y})\mathrm{d}\boldsymbol{\psi},$$
$$\mathrm{DIC}_1 = D(\bar{\boldsymbol{\psi}}) + 2P_D.$$

The likelihood $p(\mathbf{y}|\boldsymbol{\psi})$ can be obtained from,

$$p(\mathbf{y}|\boldsymbol{\psi}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\psi})\mathrm{d}\boldsymbol{\theta}.$$

Clearly, parameters in focus are different in these two cases, leading to different likelihood functions for constructing DIC.

Parameters in focus bear some similarity to parameters of interest in statistical inference for a model that involve two types of parameters, parameters of interest (say $\boldsymbol{\theta}$) and nuisance parameters (say $\boldsymbol{\psi}$). To remove the influence of nuisance parameters under the Bayesian framework, $\boldsymbol{\psi}$ is integrated out to get a marginal likelihood on $\boldsymbol{\theta}$. This marginal likelihood is often called the integrated likelihood, that is,

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}|\boldsymbol{\theta},\boldsymbol{\psi})p(\boldsymbol{\psi})\mathrm{d}\boldsymbol{\psi}.$$

Berger et al (1999) discussed the use of integrated likelihood for statistical inference.

It is useful to point out that Claeskens and Hjort (2003) also defined the concept, focus. However, their definition is different from that in the Spiegelhalter et al. (2002). In Claeskens and Hjort (2003) the focus $\phi$ is a single parameter that is defined as a real-value function of parameters $\boldsymbol{\theta}$, say $\phi = g(\boldsymbol{\theta})$. Given a loss function about the focus, such as the mean-squared error (MSE), Claeskens and Hjort (2003) proposes a focused information criterion under the frequentist framework. Clearly, the focus in Claeskens and Hjort (2003) is determined by a single parameter related to the purpose of selecting the optimal model whereas in Spiegelhalter et al. (2002) the focus refers to the parameters of interest for the purpose of determining the likelihood.

### 3.2 Decision-theoretic justification of DIC

As acknowledged in Spiegelhalter et al. (2002) (Section 7.3 on Page 603 and the first paragraph on Page 605), the justification of DIC is informal and heuristic. In this section we provide a rigorous decision-theoretic justification of DIC, in the same spirit as the justification of AIC. In our view, the lack of rigorous justification of DIC lies in the inappropriate specification of loss function. When a proper loss function is selected, DIC can be justified asymptotically.

Given that DIC is a Bayesian version of AIC, before we justify DIC, it is useful to review AIC and its decision-theoretic justification. Let the true data generating process (DGP) be $g(\mathbf{y})$, $\mathbf{y}_{rep} = (\mathbf{y}_{1,rep}, \mathbf{y}_{2,rep}, \cdots, \mathbf{y}_{n,rep})$ be the independent replicate data generated by the same mechanism that gives rise to the observed data $\mathbf{y}$. Consider a candidate parametric model, $M$, denoted by $p(\mathbf{y}|M, \boldsymbol{\theta})$ to fit the data, where $\boldsymbol{\theta}$ is the parameter with $P$ dimensions and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq R^P$. When there is no confusion, we simply write $p(\mathbf{y}|M, \boldsymbol{\theta})$ as $p(\mathbf{y}|\boldsymbol{\theta})$. In the literature, the KL divergence is used to describe the difference between two models and given by:

$$KL[p(x), q(x)] = \int p(x) \ln \frac{p(x)}{q(x)} dx.$$

The quantity that measures the quality of the candidate model in terms of its ability to make predictions is given by the KL divergence between $g(\mathbf{y}_{rep})$ and $p(\mathbf{y}_{rep}|\boldsymbol{\theta})$,

$$\begin{aligned} KL\left[g(\mathbf{y}_{rep}), p(\mathbf{y}_{rep}|\boldsymbol{\theta})\right] &= E_{\mathbf{y}_{rep}}\left[\ln \frac{g(\mathbf{y}_{rep})}{p(\mathbf{y}_{rep}|\boldsymbol{\theta})}\right] \\ &= E_{\mathbf{y}_{rep}}(\ln g(\mathbf{y}_{rep}) - E_{\mathbf{y}_{rep}}(\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})), \end{aligned} \tag{7}$$

where the expectation is about $g(\mathbf{y}_{rep})$. Since $g(\mathbf{y}_{rep})$ is the true DGP and thus $E_{\mathbf{y}_{rep}}(\ln g(\mathbf{y}_{rep}))$ is independent with the candidate models, it is dropped from the above equation. The smaller this KL divergence, the better the candidate model in predicting $g(\mathbf{y}_{rep})$.

Let $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ be the ML estimate of $\boldsymbol{\theta}$ obtained from $\mathbf{y}$. Since $\boldsymbol{\theta}$ is unknown, it is replaced by $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ and $p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y}))$ is the plug-in predictive distribution. Thus, up to a constant, $E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))$ represents the KL divergence between the predictive distribution of the candidate model and the true DGP. Although a natural estimate of $E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))$ is $-2\ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))$, it is asymptotically biased. Let

$$c(\mathbf{y}) = E_{\mathbf{y}_{rep}}\left(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right) - \left(-2\ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))\right). \tag{8}$$

Under a set of regularity conditions and the dominated convergence theorem, one can show that $E_{\mathbf{y}}(c(\mathbf{y})) \to 2P$ where the expectation is about $g(\mathbf{y})$; see for example, Burnham and Anderson (2002). Hence, if we let AIC $= -2\ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y})) + 2P$, as $n \to \infty$,

$$E_{\mathbf{y}}(\text{AIC}) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))) \to 0.$$

To see why a penalty term, $2P$, is needed in AIC, let $\boldsymbol{\theta}^t := \arg\min_\theta KL[g(\mathbf{y}), p(\mathbf{y}|\boldsymbol{\theta})]$ be the pseudo-true parameter value (Sawa, 1978; Gourieroux, et al, 1984), $\widehat{\boldsymbol{\theta}}(\mathbf{y}_{rep})$ be the ML estimate of $\boldsymbol{\theta}$ obtained from $\mathbf{y}_{rep}$. Note that

$$\begin{aligned} E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))) &= \underbrace{\left[E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}_{rep})))\right]}_{(T1)} \\ &+ \underbrace{\left[E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}_{rep})))\right]}_{(T2)} \\ &+ \underbrace{\left[E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t))\right]}_{(T3)}. \end{aligned}$$

Clearly, the term in $T1$ is the same as $E_{\mathbf{y}}(-2\ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y}))$. The term in $T2$ is the expectation of the likelihood ratio statistic based on the replicate data. Under a set of regularity conditions that ensure $\sqrt{n}$-consistency and asymptotic normality of the ML estimates and a dominated condition, $T2$ is approximately the same as the expectation of $\chi^2_{(P)}$which is $P$. To approximate the term in $T3$, if $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is a consistent estimate of $\boldsymbol{\theta}^t$, we have

$$\begin{aligned} T3 &= E_{\mathbf{y}}\left\{-2E_{\mathbf{y}_{rep}}\left[\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial\theta}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]\right\} \\ &\quad +E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left[-\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial\theta\partial\theta'}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]\right\}+o(1). \end{aligned}$$

By the definition of $\boldsymbol{\theta}^t$, $E_{\mathbf{y}_{rep}}\left[\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)/\partial\boldsymbol{\theta}\right]=0$, implying that

$$E_{\mathbf{y}}\left\{-2E_{\mathbf{y}_{rep}}\left[\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial\theta}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]\right\}=-2E_{\mathbf{y}_{rep}}\left(\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial\theta}\right)E_{\mathbf{y}}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)=0.$$

Consequently, under the same regularity conditions for approximating $T2$, we have

$$T3=\mathbf{tr}\left\{E_{\mathbf{y}_{rep}}\left[\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial\theta\partial\theta'}\right]E_{\mathbf{y}}\left[-\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]\right\}+o(1)=P+o(1),$$

where $\mathbf{tr}$ denotes the trace of a matrix.

It is clear that the decision-theoretic justification of AIC requires a careful choice of the KL loss function, the use of ML estimation and the a set of regularity conditions for ensuring $\sqrt{n}$-consistency and asymptotic normality of the ML estimates. The penalty term in AIC arises from two sources. First, the pseudo-true value has to be estimated. Second, the estimate obtained from the observed data is not the same as that from the replicate data. It is important to point out that the justification of AIC requires the true DGP be nested by the candidate model.

In practice, empirical researchers may have prior information about model parameters which may in turn reduce the model complexity. However, AIC does not work in models with informative prior information. DIC intends to take account of prior information.

To our surprise, Spiegelhalter, et al (2002) did not explicitly specify the KL function when developing DIC. However, from Equation (33) and Equation (40) in their paper and the loss function defined in the first paragraph on Page 603, namely, $-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$, one can deduce that the following KL function

$$KL\left[p(\mathbf{y}_{rep}|\boldsymbol{\theta}), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] = E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left[\ln\frac{p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)}\right], \tag{9}$$

was used where $\bar{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}$ for a candidate model. Hence,

$$KL\left[p(\mathbf{y}_{rep}|\boldsymbol{\theta}), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] = E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})\right) - E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right). \tag{10}$$

In Equation (33), Spiegelhalter, et al (2002) dealt with $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right)$ directly and ignored the first term in the right hand side of Equation (10). On Page 604, they argued that, if

$$c\left(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}(\mathbf{y})\right) := E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left[\left(-2\ln p(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right) - \left(-2\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}(\mathbf{y}))\right)\right],$$

then

$$\int\left\{E_{\boldsymbol{\theta}|\mathbf{y}}\left[c\left(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] - 2P_D\right\}p(\mathbf{y})d\mathbf{y} \to \mathbf{0}, \tag{11}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. This leads to $\mathrm{DIC}_1 = D(\bar{\boldsymbol{\theta}}) + 2P_D$. The convergence in (11) was proved without any conditions being specified. Clearly, an implicit assumption made in this heuristic argument is that the first term in the right hand side of Equation (10) is constant across candidate models and thus dropped from (10). While the treatment mimics Equation (7) in the development of AIC, unfortunately, one cannot claim that $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})\right)$ is the same across all candidate models. This is because, as Spiegelhalter, et al (2002) said in the second paragraph on Page 604, "we are taking a Bayesian perspective" and "we replace the pseudo-true value by a random quantity". As a result, $\boldsymbol{\theta}$ in the first term in the right hand side of Equation (10) is model dependent and in general $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})\right)$ takes a different value for each candidate model.

From the discussion above, clearly $KL\left[p(\mathbf{y}_{rep}|\boldsymbol{\theta}), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ is not the proper KL loss function to justify DIC. A new KL loss function is needed. To do so, let $p(\mathbf{y}_{rep}|\mathbf{y})$ be the Bayesian predictive distribution where $p(\mathbf{y}_{rep}|\mathbf{y}) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}$. By mimicking the development of AIC, we propose the following KL loss function based on the Bayesian predictive distribution,

$$KL\left[g(\mathbf{y}_{rep}), p(\mathbf{y}_{rep}|\mathbf{y})\right] = E_{\mathbf{y}_{rep}}\left(\ln g\left(\mathbf{y}_{rep}\right)\right) - E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right). \tag{12}$$

A better model is expected to yield a smaller value for the KL function. Since $g(\mathbf{y}_{rep})$ is the true DGP and $\mathbf{y}_{rep}$ is an independent replication as in AIC, $E_{\mathbf{y}_{rep}}\left(\ln g\left(\mathbf{y}_{rep}\right)\right)$ is model-independent. Therefore, it is the same across all candidate models and can be dropped from

(12) when comparing models. As a result, we propose to choose a model that gives the smallest value of

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right)=\int\int -2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)g(\mathbf{y}_{rep})g(\mathbf{y})\mathrm{d}\mathbf{y}_{rep}\mathrm{d}\mathbf{y},$$

which is equivalent to minimizing the following expectation of the KL loss function

$$E_{\mathbf{y}}KL\left[g(\mathbf{y}_{rep}),p(\mathbf{y}_{rep}|\mathbf{y})\right]=\int\int KL\left[g(\mathbf{y}_{rep}),p(\mathbf{y}_{rep}|\mathbf{y})\right]g(\mathbf{y})\mathrm{d}\mathbf{y}.$$

We are now in the position to provide a rigorous decision-theoretic justification of $\mathrm{DIC}_1$ based on a set of regularity conditions. Let $L_n(\boldsymbol{\theta})=\ln p(\boldsymbol{\theta}|\mathbf{y})$, $L_n^{(1)}(\boldsymbol{\theta})=\partial\ln p(\boldsymbol{\theta}|\mathbf{y})/\partial\boldsymbol{\theta}$, $L_n^{(2)}(\boldsymbol{\theta})=\partial^2\ln p(\boldsymbol{\theta}|\mathbf{y})/\partial\boldsymbol{\theta}\boldsymbol{\theta}'$. In this paper, we impose the following regularity conditions.

**Assumption 1**: There exists a finite sample size $n^*$, for $n>n^*$, there is a local maximum at $\hat{\boldsymbol{\theta}}_m$ so that $L_n^{(1)}\left(\hat{\boldsymbol{\theta}}_m\right)=0$ and $L_n^{(2)}\left(\hat{\boldsymbol{\theta}}_m\right)$ is a negative definite matrix. Obviously, $\hat{\boldsymbol{\theta}}_m$ is the posterior mode and $L_n^{(2)}(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})/n=O_p(1)$.

**Assumption 2**: The largest eigenvalue of $\left[-L_n^{(2)}(\hat{\boldsymbol{\theta}}_m)\right]^{-1}$, $\sigma_n^2$, goes to zero when $n\to\infty$.

**Assumption 3**: For any $\epsilon>0$, there exists an integer $n^{**}$ and some $\delta>0$ such that for any $n>\max\{n^*,n^{**}\}$ and $\boldsymbol{\theta}\in H\left(\hat{\boldsymbol{\theta}}_m,\delta\right)=\left\{\boldsymbol{\theta}:||\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m||\leq\delta\right\}$, $L_n^{(2)}(\boldsymbol{\theta})$ exists and satisfies

$$-A(\epsilon)\leq L_n^{(2)}(\boldsymbol{\theta})L_n^{-(2)}\left(\hat{\boldsymbol{\theta}}_m\right)-\mathbf{I}_P\leq A(\epsilon),$$

where $\mathbf{I}_P$ is a $P\times P$ identity matrix, $A(\epsilon)$ a $P\times P$ positive semi-definite symmetric matrix whose largest eigenvalue goes to zero as $\epsilon\to 0$. $A\leq B$ means that $A_{ij}\leq B_{ij}$ for all $i,j$.

**Assumption 4**: For any $\delta>0$, as $n\to\infty$,

$$\int_{\boldsymbol{\Theta}-H(\hat{\boldsymbol{\theta}}_m,\delta)}p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}\to 0,$$

where $\boldsymbol{\Theta}$ is the support of $\boldsymbol{\theta}$.

**Assumption 5**: For any two $\theta_i,\theta_j,i,j=1,2,\cdots,P$, we have

$$\int|\theta_i|\,p(\theta_i|\mathbf{y})d\theta_i<\infty,\int|\theta_i\theta_j|\,p(\theta_i,\theta_j|\mathbf{y})d\theta_i d\theta_j<\infty.$$

**Assumption 6**: Assume the standard ML theory, such as $\sqrt{n}$-consistency, asymptotic normality, and asymptotic efficiency, is applicable. Furthermore, for any replicate data $\mathbf{y}_{rep}$, let the Hessian information matrix, $\mathbf{H}(\boldsymbol{\theta})$, and Fisher information matrix, $\mathbf{J}(\boldsymbol{\theta})$, be

$$\begin{aligned}\mathbf{H}(\boldsymbol{\theta})&=\int-\frac{1}{n}\left[\frac{\partial^2\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]g(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}_{rep}=0,\\ \mathbf{J}(\boldsymbol{\theta})&=E\left\{\left[\frac{1}{\sqrt{n}}\frac{\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]\left[\frac{1}{\sqrt{n}}\frac{\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\right]'\right\}\\ &=\int\frac{1}{n}\frac{\partial\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})'}{\partial\boldsymbol{\theta}'}g(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}.\end{aligned}$$

For any null sequence $k_n$, it is assumed that

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_t||\leq k_n} \left\| -\frac{1}{n}\left[\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] - \mathbf{H}(\boldsymbol{\theta}_t)\right\| \xrightarrow{p} 0.$$

$$\sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_t||\leq k_n} \left\| \frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}' - \mathbf{J}(\boldsymbol{\theta}_t)\right\| \xrightarrow{p} 0.$$

**Lemma 3.1** *Under Assumptions 1-5, conditional on the observed data* $\mathbf{y}$*, we have*

$$\bar{\boldsymbol{\theta}} = E\left[\boldsymbol{\theta}|\mathbf{y}\right] = \hat{\boldsymbol{\theta}}_{\boldsymbol{m}} + o_p(n^{-1/2}),$$
$$V\left(\hat{\boldsymbol{\theta}}_m\right) = E\left[\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m\right)\left(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m\right)' |\mathbf{y}\right] = -L_n^{-(2)}\left(\hat{\boldsymbol{\theta}}_m\right) + o_p(n^{-1}).$$

**Remark 3.1** *Assumptions 1-4 have been used in the literature to develop the standard Bayesian large sample theory for dynamic models and non-dynamic models; see, for example, Chen (1985), Kim (1994, 1998), Geweke (2005). Under the different regularity conditions, the Bernstein-von Mises theorem shows that the posterior distribution converges to a normal distribution with the ML estimator as its mean and the inverse of the second derivative of the log-likelihood evaluated at the ML estimator as its covariance. Based on Bernstein-von Mises theorem, Ghosh and Ramamoorthi (2003) developed the same results in Lemma 3.1 for the iid case. Under Assumptions 1-4, Chen (1985) shows that the posterior distribution converges to a normal distribution with the posterior mode as its mean and the inverse of the second derivative of the log-posterior evaluated at mode as its covariance. Based on the results of Chen (1985), in Lemma 3.1 we extend the results of Ghosh and Ramamoorthi (2003) to more general cases. Assumption 6 is useful to establish the equivalence of* $\mathbf{H}(\boldsymbol{\theta}_0)$ *and* $\mathbf{J}(\boldsymbol{\theta}_0)$*.*

**Theorem 3.1** *Under Assumptions 1-6, when the prior* $p(\boldsymbol{\theta}) = O_p(1)$*, it can be shown that,*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] = E_{\mathbf{y}}\left[DIC_1 + o_p(1)\right] = E_{\mathbf{y}}\left[D(\bar{\boldsymbol{\theta}}) + 2P_D + o_p(1)\right],$$

*where* $P_D$ *is defined in (4).*

**Remark 3.2** *If, in addition, there exists* $Z(\mathbf{y})$ *such that* $E_{\mathbf{y}}[Z(\mathbf{y})] < \infty$ *and*

$$\left|E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] - DIC_1\right| \leq Z(\mathbf{y}),$$

*for all* $n$*, the dominated convergence theorem holds. In this case we have,*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] = E_{\mathbf{y}}\left[DIC_1 + o_p(1)\right] = E_{\mathbf{y}}\left[DIC_1\right] + o(1).$$

*Hence,* $DIC_1$ *is an asymptotic unbiased estimator of the proposed KL loss function.*

The asymptotic justification of $\text{DIC}_1$ requires that the candidate model nest the true model, and that the posterior distribution is approximately normal with the posterior mean converging to posterior mode and the posterior variance converging to zero. These requirements parallel to those in AIC where the candidate models nest the true model and the ML estimator is $\sqrt{n}$-consistent and asymptotically normally distributed. To see the importance of the asymptotic normality, Spiegelhalter, et al (2002) show that, when the prior is noninformative, $P_D$ is approximately the same as the number of parameters, $P$**.** In this case $\text{DIC}_1$ is the Bayesian version of AIC.

In AIC, the degrees of freedom are used to measure the model complexity. In the Bayesian framework, the prior information often imposes additional restrictions on the parameter space and hence the degrees of freedom may be reduced by the prior information. In this case, $P_D$ may not be close to $P$. A useful contribution of $\text{DIC}_1$ is to provide a way to measure the model complexity when the prior information is incorporated; see Brooks (2002).

If $p(\mathbf{y}|\boldsymbol{\theta})$ has a closed-form expression, $\text{DIC}_1$ is trivially computable from the MCMC output. The computational tractability, together with the versatility of MCMC and the fact that $\text{DIC}_1$ is incorporated into a Bayesian software, WinBUGS, allows $\text{DIC}_1$ to enjoy a very wide range of applications.[2] However, if $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form, such as in random effects models and state space models, computing $\text{DIC}_1$ may become infeasible, or at least, very time consuming.

# 4 Bayesian Comparison of Latent Variable Models

## 4.1 DIC for latent variable models

As described in Section 2, in latent variable models, there are three types of variables, the observed data $\mathbf{y}$, the latent variables $\mathbf{z}$, and the parameters $\boldsymbol{\theta}$. In the frequentist framework, the likelihood, $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\text{d}\mathbf{z}$, is clearly defined. In this case, only $\boldsymbol{\theta}$ is treated as parameters and there is no confusion in defining AIC. In the Bayesian framework, however, depending on whether the latent variables $\mathbf{z}$ are treated as parameters or variables, three likelihood functions may be used,

$$p(\mathbf{y}|\boldsymbol{\theta}),\ p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}),\ \text{and}\ p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}),$$

termed as the observed-data, complete-data, conditional likelihood functions, respectively. Obviously, $\text{DIC}_1$ is based on the observed-data likelihood function, which is computationally demanding for many latent variable models. With these three likelihood functions, Celeux et al (2006) considered and compared eight versions of DIC. Based on the observed-data

[2] According to, Spiegelhalter et al. (2014), Spiegelhalter et al. (2002) was the third most cited paper in international mathematical sciences between 1998 and 2008, and up to November 2013 it had over 2500 citations on the Web of Knowledge and over 4600 on Google Scholar.

likelihood $p(\mathbf{y}|\boldsymbol{\theta})$, the first category includes

$$\begin{aligned}
\mathrm{DIC}_1 &= -4E_{\boldsymbol{\theta}|\mathbf{y}}\left[\ln p(\mathbf{y}|\boldsymbol{\theta})\right] + 2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right), \\
\mathrm{DIC}_2 &= -4E_{\boldsymbol{\theta}|\mathbf{y}}\left[\ln p(\mathbf{y}|\boldsymbol{\theta})\right] + 2\ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right), \\
\mathrm{DIC}_3 &= -4E_{\boldsymbol{\theta}|\mathbf{y}}\left[\ln p(\mathbf{y}|\boldsymbol{\theta})\right] + 2\ln\left\{E_{\boldsymbol{\theta}|\mathbf{y}}\left[p(\mathbf{y}|\boldsymbol{\theta})\right]\right\},
\end{aligned}$$

where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mode.

Based on the complete-data likelihood $p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})$, the second category includes

$$\begin{aligned}
\mathrm{DIC}_4 &= -4E_{\boldsymbol{\theta},\mathbf{z}|\mathbf{y}}[\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})] + 2E_{\mathbf{z}|\mathbf{y}}\ln p\left(\mathbf{y},\mathbf{z}|E_{\theta|\mathbf{y},\mathbf{z}}\left[\boldsymbol{\theta}|\mathbf{y},\mathbf{z}\right]\right), \\
\mathrm{DIC}_5 &= -4E_{\boldsymbol{\theta},\mathbf{z}|\mathbf{y}}\left[\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})\right] + 2\ln p\left(\mathbf{y},\hat{\mathbf{z}}(\mathbf{y})|\hat{\boldsymbol{\theta}}(\mathbf{y})\right), \\
\mathrm{DIC}_6 &= -4E_{\boldsymbol{\theta},\mathbf{z}|\mathbf{y}}\left[\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})\right] + 2E_{\mathbf{z}|\mathbf{y},\hat{\boldsymbol{\theta}}(\mathbf{y})}\left[\ln p\left(\mathbf{y},\mathbf{z}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right],
\end{aligned}$$

where in $\mathrm{DIC}_5$, $\hat{\mathbf{z}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y})$ are the joint Bayesian estimators, such as the joint maximum a posteriori (MAP) estimators of $(\mathbf{z},\boldsymbol{\theta})$; in $\mathrm{DIC}_6$, $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is an estimator of $\boldsymbol{\theta}$ based on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$.

Based on the conditional likelihood $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$, the third category includes

$$\begin{aligned}
\mathrm{DIC}_7 &= -4E_{\boldsymbol{\theta},\mathbf{z}|\mathbf{y}}\left[\ln p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})\right] + 2\ln p\left(\mathbf{y}|\hat{\mathbf{z}}(\mathbf{y}),\hat{\boldsymbol{\theta}}(\mathbf{y})\right), \\
\mathrm{DIC}_8 &= -4E_{\boldsymbol{\theta},\mathbf{z}|\mathbf{y}}\left[\ln p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})\right] + 2E_{\mathbf{z}|\mathbf{y}}\left[\ln p\left(\mathbf{y}|\mathbf{z},\hat{\boldsymbol{\theta}}(\mathbf{y},\mathbf{z})\right)\right],
\end{aligned}$$

where in $\mathrm{DIC}_7$, $\mathbf{z}$ is treated as parameters so that $\hat{\mathbf{z}}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}(\mathbf{y})$ are the joint Bayesian estimator, such as the posterior mean or the MAP estimator of $(\mathbf{z},\boldsymbol{\theta})$; in $\mathrm{DIC}_8$, $\hat{\boldsymbol{\theta}}(\mathbf{y},\mathbf{z})$ is an estimator of $\boldsymbol{\theta}$ based on $p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})$.

When constructing DIC, one needs to define parameters in focus. If $\boldsymbol{\theta}$ is the parameters in focus, the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is used to construct DIC. This choice of focus leads to $\mathrm{DIC}_1$ and $\mathrm{DIC}_2$. If the latent variables $\mathbf{z}$ and the parameters $\boldsymbol{\theta}$ are in focus, the conditional likelihood $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ is used for constructing DIC. This choice of focus leads to $\mathrm{DIC}_7$ and $\mathrm{DIC}_8$. Clearly, the other three versions, $\mathrm{DIC}_4$, $\mathrm{DIC}_5$ and $\mathrm{DIC}_6$, are logically incoherent as far as the focus is concerned. This is because the latent variables $\mathbf{z}$ are treated as both variables and parameters. Similarly, $\mathrm{DIC}_8$ is logically incoherent because parameters in focus are $(\mathbf{z},\boldsymbol{\theta})$ in the first term, but they are $\mathbf{z}$ in the second term. As pointed out by Plummer (2006), $\mathrm{DIC}_3$ does not have an unambiguous focus corresponding to it and it is not clear which likelihood is used to construct $\mathrm{DIC}_3$. Therefore, only $\mathrm{DIC}_1$, $\mathrm{DIC}_2$, and $\mathrm{DIC}_7$ are logically coherent. Although Celeux et al (2006) recommended $\mathrm{DIC}_3$ and $\mathrm{DIC}_4$ based on a real example, neither $\mathrm{DIC}_3$ nor $\mathrm{DIC}_4$ is logically coherent.

Celeux et al (2006) compared $\mathrm{DIC}_1$ with $\mathrm{DIC}_2$ and found the evidence that $\mathrm{DIC}_2$ is better than $\mathrm{DIC}_1$ since the posterior mode can ensure that $P_D$ is positive, but the posterior mean

cannot. However, $\text{DIC}_1$ and $\text{DIC}_2$ are asymptotically equivalent following Lemma 3.1. In practice, the posterior mode is more difficult to compute than the posterior mean.

For many latent variable models, such as state-space models, including linear Gaussian state space models, the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form.[3] In this case, both $\text{DIC}_1$ and $\text{DIC}_2$ are very difficult to compute because it needs to evaluate the observed-data likelihood at each MCMC iteration.

$\text{DIC}_1$ is monitored and reported in WinBUGS when there is no latent variable. To compute $\text{DIC}_1$, it is generally required the observed likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ be available in closed-form because we need to evaluate $E_{\boldsymbol{\theta}|\mathbf{y}}\left[\ln p(\mathbf{y}|\boldsymbol{\theta})\right] \approx \frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(m)}\right)$. Given that $M$ is usually very large, computing $\frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(m)}\right)$ without knowing the analytical form of $\ln p(\mathbf{y}|\boldsymbol{\theta}))$ is very costly. In $\text{DIC}_7$, the latent variables are regarded as parameters and $\ln p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ has an analytical expression. Hence, it is easy to compute $\frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{y}|\mathbf{z}^{(m)},\boldsymbol{\theta}^{(m)}\right)$. That is why, when there are latent variables, $\text{DIC}_7$ is monitored and reported in WinBUGS, following the suggestion of Spiegelhalter, et al (2002). Clearly the use of $\text{DIC}_7$ is for computational convenience, as explained in Spiegelhalter, et al (2002).

### 4.2 RDIC for latent variable models

From a theoretical viewpoint, $\text{DIC}_7$ has a few problems. Firstly, due to the data augmentation, the number of the latent variables often increases with the sample size in latent variable models. This may lead to the well-known incidental parameter problem where information about these incidental parameters stops accumulating after a finite number of observations, often one, have been taken; see for example Neyman and Scott (1948) and Lancaster (2000). A consequence of the incidental parameter problem is that the ML estimator is inconsistent. Similarly, the Bayesian large sample theory becomes invalid; see Page 89-90 of Gelman, et al (2013) for examples. The failure of the standard asymptotic theory invalidates the asymptotic justification of DIC. Secondly, if the latent variable can be treated as parameters, an incoherent inference problem will result. That is, when one model can be rewritten as distributional representation of another model with latent variables and the same prior is used in the two models, the different DIC values can be obtained. A simple example is the student-t distribution which can be rewritten as a normal-gamma scale mixture representation. In Section 8.2 of Spiegelhalter, et al (2002), Models 4 and 5 are predictively identical but their DIC values are quite different. The same difficulty also shows up in Model 8 of Berg, et al (2004). Thirdly, when the latent variables are discrete, such as component indicators in Markov switching models, generally, Bayesian estimator is not a discrete value which can cause some logic problems. Fourthly, due to the data augmentation, the dimension of the pa-

[3] For linear Gaussian state space models, to do ML, the Kalman filter can be used to obtain the likelihood function numerically. Numerically more efficient algorithms have been developed in the recent literature; see for example, Chan and Jeliazkov (2009).

rameter space becomes larger and hence we expect $\text{DIC}_7$ be very sensitive to transformations of latent variables.

To illustrate the first problem, consider the following example: Let $y_i|\alpha_i, \sigma^2 \sim N(\alpha_i, \sigma^2)$, $\alpha_i \sim N(0,1)$ for $i = 1, ..., n$. Clearly $y_i|\sigma^2 \sim N(0, \sigma^2 + 1)$ and thus the ML estimate of $\sigma^2$ is $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} y_i^2 - 1$. It is straightforward to show $\widehat{\sigma}^2$ is $\sqrt{n}$-consistent and asymptotically normal. However, if $\{\alpha_i\}_{i=1}^n$ are treated as parameters, they are incidental in the sense of Neyman and Scott (1948). As a result, the incidental parameter problem arises. The ML estimate of $\alpha_i$ is $\widehat{\alpha}_i = y_i \sim N(\alpha_i, \sigma^2)$. So $\widehat{\alpha}_i$ is unbiased but inconsistent. Similarly, in the case when $\sigma^2 = 1$, the posterior distribution $\alpha_i|y_i \sim N(0.5y_i, 0.5)$. The posterior mean (which is also the posterior mode) is not close to the ML estimate and the posterior variance does not go to zero as $n$ grows. Both the standard ML theory and Bayesian large sample theory fail to hold. These results are not surprising as only one observation ($y_i$) contains information about $\alpha_i$. Lancaster (2000) surveys the problem in the statistics and econometrics literature.

To illustrate the last problem, we consider a simple transformation of latent variables in the well-known Clark model ( Clark, 1973) which is given by,

$$\text{Model 1} : y_t \sim N(\mu, \exp(h_t)), h_t \sim N(0, \sigma^2), t = 1, \cdots, n. \tag{13}$$

An equivalent representation of the model is

$$\text{Model 2} : y_t \sim N(\mu, \sigma_t^2), \sigma_t^2 \sim LN(0, \sigma^2), t = 1, \cdots, n, \tag{14}$$

where $LN$ denotes the log-normal distribution. In Model 2 the latent variable is the volatility $\sigma_t^2$ while the latent variable is the logarithmic volatility $h_t = \ln \sigma_t^2$ in Model 1. Suppose the parameters of interest are $\mu$ and $\sigma^2$. With the same focus, the two models are identical and hence are expected to have the same DIC and $P_D$. To calculate the $P_D$ component in $\text{DIC}_7$, we simulate 1000 observations from the model with $\mu = 0, \sigma^2 = 0.5$. Vague priors are selected for the two parameters, namely, $\mu \sim N(0, 100)$, $\sigma^{-2} \sim \Gamma(0.001, 0.001)$. We run Gibbs sampler to make 240,000 simulated draws from the posterior distributions. The first 40,000 are discarded as burn-in samples. The remaining observations with every 10th observation are collected as effective observations for statistical inference. With the data augmentation, the latent variables, $h_t$ and $\sigma_t^2$ are regarded as parameters, and we find that $P_D = 89.806$ for Model 1 but $P_D = 59.366$ for Model 2. The difference is very large. Given that we have the identical models and priors, and use the same dataset, the vast difference suggests that $\text{DIC}_7$ and the corresponding $P_D$ are very sensitive to transformations of latent variables.

To summarize the problems with DIC in the context of latent variable models, while $\text{DIC}_7$ is trivial to calculate and has been used widely in practice but does not have a decision-theoretic justification, $\text{DIC}_1$ is theoretically justified but infeasible to compute from the MCMC output since $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form.

In this section we introduce a robust version of DIC, denoted as RDIC, to approximate $\text{DIC}_1$ and then show how to compute RDIC from the MCMC output. RDIC is defined as

$$\text{RDIC} = D(\bar{\boldsymbol{\theta}}) + 2\mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\} = D(\bar{\boldsymbol{\theta}}) + 2P_D^*, \tag{15}$$

where

$$P_D^* = \mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}, \tag{16}$$

and

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

Interestingly, in Equation (15) on Page 590, Spiegelhalter, et al. (2002) obtained the expression for $P_D^*$ and claimed that $P_D^*$ approximates the $P_D$ component in $\text{DIC}_1$. Unfortunately, to the best of our knowledge, $P_D^*$ has never been implemented in practice and WinBUGS does not report $P_D^*$. Moreover, the conditions under which $P_D^* \approx P_D$ holds true were not specified in Spiegelhalter, et al (2002). The order of the approximation error is unknown. To justify the choice of RDIC, we will show that RDIC approximates $\text{DIC}_1$ and $P_D^*$ approximates the $P_D$ component in $\text{DIC}_1$ and obtain the order for the approximation error.

**Theorem 4.1** *Under Assumptions 1-6, assume the prior $p(\boldsymbol{\theta}) = O_p(1)$, it can be shown that,*

$$P_D^* = P_D + o_p(1), \;\; DIC_1 = RDIC + o_p(1),$$

**Corollary 4.2** *Assume the prior $p(\boldsymbol{\theta}) = O_p(1)$. It can be shown that*

$$P_D^* = P_D + o_p(1) = P + o_p(1), \;\; RDIC = DIC_1 + o_p(1) = AIC + o_p(1).$$

*and*

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right) = E_{\mathbf{y}}\left[RDIC + o_p(1)\right].$$

Theorem 4.1 extends the result in Equation (15) of Spiegelhalter, et al (2002) by specifying the conditions under which $P_D$ approximates $P_D^*$ and $P$, and $\text{DIC}_1$ approximates RDIC. Corollary 4.2 shows that the order of difference between AIC and $\text{RDIC}/\text{DIC}_1$ is $o_p(1)$. For this reason, both RDIC and $\text{DIC}_1$ can be regarded as the Bayesian version of AIC. Furthermore, Corollary 4.2 justifies RDIC by showing that RDIC is asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right)$ if the dominated convergence theorem holds. As $\text{DIC}_1$, RDIC is defined from the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ where the focus is on $\boldsymbol{\theta}$ only. Unlike $\text{DIC}_7$, the latent variables are not parameters in focus in RDIC.

To understand how the prior information can affect $P_D^*$ in finite sample, a higher order approximation than what has been stated in Theorem 4.1 and Corollary 4.2 is needed. Following the literature on the finite sample theory, such as in Rilstone, Srivatsava and Ullah (1996), Bester and Hansen (2008) and Li et al. (2015), let $\nabla^j$ denote the $jth$ derivative,

$\psi_n(\boldsymbol{\theta}) = \frac{1}{n}\nabla_{\boldsymbol{\theta}} \ln p(\mathbf{y}|\boldsymbol{\theta})$, $\gamma^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \ln p(\boldsymbol{\theta})$. Let the bar over a function indicate its expectation so that $\overline{A}(\boldsymbol{\theta}) = E[A(\boldsymbol{\theta})]$. Further let $H_j(\boldsymbol{\theta}) = \nabla^j \psi_n(\boldsymbol{\theta})$, $Q(\boldsymbol{\theta}) = \overline{H}_1^{-1}(\boldsymbol{\theta})$, $\varrho_{-1}(\boldsymbol{\theta}) = \frac{1}{2n}\left[[vec(Q(\boldsymbol{\theta}))]' \otimes Q(\boldsymbol{\theta})\right] vec\left(\overline{H}_2(\boldsymbol{\theta})\right)$ and $C(\boldsymbol{\theta})$ be a continuous bounded function of $\boldsymbol{\theta}$. For notational convenience, we suppress the argument of a function when it is evaluated at $\boldsymbol{\theta}_0$. A higher order approximation for $P_D^*$ is given in the following Corollary.

**Corollary 4.3** *Under Assumption 1-6 and other regularity conditions as stated in Li et al. (2015), if the prior $p(\boldsymbol{\theta}) = O_p(1)$, we have*

$$P_D^* = P - \frac{1}{n} tr\left[\nabla\gamma^{\boldsymbol{\theta}} Q\right] - \frac{1}{n} tr\left[Q^{-1}\overline{C}\right] - tr\left[\overline{H}_2\left(I_P \otimes \varrho_{-1}\right) Q\right] + o_p\left(n^{-1}\right), \tag{17}$$

*where $\varrho_{-1}$ has the order $O_p(n^{-1})$.*

The second term in the right hand side of Equation (17) explicitly depends on $\nabla\gamma^{\boldsymbol{\theta}}$, the second order derivative of the prior. The third term and the fourth term are dependent on the likelihood function evaluated at $\boldsymbol{\theta}_0$ but independent on the prior. All these three terms are of order $O_p(n^{-1})$. Hence, the effect of the prior on $P_D^*$ is through $\nabla\gamma^{\boldsymbol{\theta}}$ and has the order of $O_p(n^{-1})$. Equation (17) shows how $P_D^*$ incorporates prior information in finite sample.

For latent variable models, while the number of model parameters ($P$) is fixed, the number of latent variables may increase as the sample size increases. In the definition of RDIC, the latent variables are not regarded as parameters. Consequently, there is no incidental parameter problem. Also, the problem of parameter transformation is less serious. For example, in the Clark model, with the same setting as before, we get $P_D^* = 1.75$ for Model 1 and $P_D^* = 1.80$ for Model 2. There is no significant difference between them. Moreover, these two values are close to 2, that is the actual number of parameters. This is what we expected given that the vague priors are used and hence $P_D^* \approx P = 2$. The small difference between $P_D^*$ and $P$ arises due to the simulation error and the priors.

To compute $P_D^*$ we only need to evaluate $\ln p(\mathbf{y}|\boldsymbol{\theta})$ once and $\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'$ once. This is in sharp contrast to $P_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}})$ where the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ needs to be evaluated $M$ times with $M$ being the number of MCMC iterations.

Like AIC, both $\text{DIC}_1$ and RDIC require the true model be nested by the candidate model. This is of course a strong assumption. Under the iid case, Ando (2010) relaxed this assumption and obtained a predictive likelihood information criterion (BPIC) that minimizes the loss function $\eta = E_{\mathbf{y}} E_{y_f}[-\ln p(y_f|\mathbf{y})]$ where $p(y_f|\mathbf{y}) = \int p(y_f|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}$ is the predictive distribution, $y_f$ is some future value. The estimator of $\eta$ is given by

$$\hat{\eta} = -\frac{1}{n} \ln p(\mathbf{y}|\mathbf{y}) + \frac{1}{2n} \mathbf{tr}\left[I^{-1}(\hat{\boldsymbol{\theta}}) J(\hat{\boldsymbol{\theta}})\right],$$

where $I(\boldsymbol{\theta})$ and $J(\boldsymbol{\theta})$ are the Hessian matrix and the Fisher information matrix. In Ando (2007), under the iid case, another BPIC was given as

$$BPIC = -2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}) + 2\mathbf{tr}[I^{-1}(\hat{\boldsymbol{\theta}}) J(\hat{\boldsymbol{\theta}})] + P.$$

Ando (2007) showed that BPIC is an estimator of the loss function

$$nE_{\mathbf{y}}E_{y_f}\left[-2\int \ln p(y_f|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right].$$

Like TIC of Takeuchi (1976), these two information criteria involve the inverse of Hessian matrix which is numerically challenging when the dimension of the parameter space is large. This is one of the reasons why TIC has not been widely used in practice. Furthermore, the derivation of these two information criteria requires the data be iid. For data in economics and finance, this requirement is often too restrictive. In addition, for many latent variable models, the ML estimator, the Hessian matrix and the Fisher information matrix are difficult to obtain.

For unit root models, Kim (1994, 1998) showed that the asymptotic normality of posterior distribution can be established under Assumptions 1-4. Hence, Lemma 3.1 holds true for unit root models. However, to develop Theorem 4.1, the standard ML asymptotic theory is required. Hence, Theorem 4.1 may not be applicable to models with a unit root or an explosive root. Based on asymptotic arguments, Phillips (1996) and Phillips and Ploberger (1996) have proposed model selection criteria for models without latent variables.

### 4.3 Computing RDIC by the EM algorithm

The definition of RDIC clearly requires the evaluation of observed-data likelihood at the posterior mean, $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$, as well as the 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. In this section we show how the EM algorithm may be used to evaluate $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$, the second derivative of the observed-data likelihood function, and hence RDIC for the latent variable models. It is important to point out that we do not need to numerically optimize any function here as in the EM algorithm for computing the ML estimates. Consequently, our method is not subject to the instability problem found in the $M$-step.

As argued in Section 2.1, the main idea of EM algorithm is to replace the observed-data log-likelihood $\ln p(\mathbf{y}|\boldsymbol{\theta})$ with the complete-data log-likelihood $\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})$. Note that

$$\ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta}) = \ln p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) + \ln p(\mathbf{y}|\boldsymbol{\theta}).$$

For any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^*$in $\Theta$, it was shown in Dempster, et al (1977) that

$$\int \ln p(\mathbf{y},\mathbf{z}|\boldsymbol{\theta})p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^*)\mathrm{d}\mathbf{z} = \int \ln p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^*)\mathrm{d}\mathbf{z} + \ln p(\mathbf{y}|\boldsymbol{\theta}),$$

and that

$$\mathcal{L}_o(\mathbf{y},\boldsymbol{\theta}) = \mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^*\right) - \mathcal{H}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^*\right), \tag{18}$$

where $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^*) = \int \ln p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^*) d\mathbf{z}$ is the so-called $\mathcal{H}$ function, $\mathcal{Q}$ is defined in Equation (2).

Following Equation (18), the Bayesian plug-in model fit, $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})$, may be obtained as

$$\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) = \mathcal{Q}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}) - \mathcal{H}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}). \tag{19}$$

It can be seen that even when $\mathcal{Q}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}})$ is not available in closed-form, it is easy to evaluate from the MCMC output because

$$\mathcal{Q}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}) = \int \ln p(\mathbf{y},\mathbf{z}|\bar{\boldsymbol{\theta}}) p(\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}}) \mathrm{d}\mathbf{z} \approx \frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{y},\mathbf{z}^{(m)}|\bar{\boldsymbol{\theta}}\right).$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ are random observations drawn from the posterior distribution $p(\boldsymbol{z}|\mathbf{y},\bar{\boldsymbol{\theta}})$.

For the second term in (19), if $p(\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}})$ is a standard distribution, $\mathcal{H}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}})$ can be easily evaluated from the MCMC output as

$$\mathcal{H}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}) = \int \ln p(\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}}) p(\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}}) d\mathbf{z} \approx \frac{1}{M}\sum_{m=1}^{M} \ln p\left(\mathbf{z}^{(m)}|\mathbf{y},\bar{\boldsymbol{\theta}}\right).$$

However, if $p(\mathbf{z}|\mathbf{y},\bar{\boldsymbol{\theta}})$ is not a standard distribution, an alternative approach has to be used, depending on the specific model in consideration. We now consider two situations.

First, if the complete-data $(\mathbf{y}_i, \mathbf{z}_i)$ are independent with $i \neq j$, and $\mathbf{z}_i$ is of low-dimension, say $\leq 5$, then a nonparametric approach may be used to approximate the posterior distribution $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$. Note that

$$\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}) = \int \ln p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) \pi(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) d\mathbf{z} = \sum_{i=1}^{n} \int \ln p(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta}) \pi(\mathbf{z}_i|\mathbf{y},\boldsymbol{\theta}) d\mathbf{z}_i = \sum_{i=1}^{n} \mathcal{H}_i(\boldsymbol{\theta}|\boldsymbol{\theta}).$$

The computation of $\mathcal{H}_i(\boldsymbol{\theta}|\boldsymbol{\theta})$ requires an analytic approximation to $p(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta})$ which can be constructed using a nonparametric method. In particular, MCMC allows one to draw some effective samples from $p(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta})$. Using these random samples, one can then use nonparametric techniques such as the kernel-based methods to approximate $p(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta})$. In a recent study, Ibrahim, et al (2008) suggested using a truncated Hermite expansion to approximate $p(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta})$.

As a simple illustration, we apply this method to the Clark model. When the Gaussian kernel method is used, we get $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) = -1448.97$, RDIC= 2901.46 for Model 1 and $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) = -1449.41$, RDIC= 2902.42 for Model 2. These two sets of numbers are nearly identical. However, if the latent variable models are regarded as parameters, we get $\mathrm{DIC}_7 = 2884.37$ for Model 1 and $\mathrm{DIC}_7 = 2852.85$ for Model 2. The highly distinctive difference between them suggests that $\mathrm{DIC}_7$ is not a reliable model selection criterion. Note that $\mathrm{DIC}_1$ is not really feasible to compute in this case.

Second, for some latent variable models, the latent variables $\mathbf{z}$ follow a multivariate normal distribution and the observed variables $\mathbf{y}$ are independent, conditional on $\mathbf{z}$. This class of models is referred to as the Gaussian latent variable models in the literature. In economics and finance, many latent variable models belong to this class, including dynamic linear models, dynamic factor models, various forms of stochastic volatility models, and credit risk models. In these models, the observed-data likelihood is non-Gaussian but has a Gaussian flavor in the sense that the posterior distribution, $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, can be expressed as,

$$p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\mathbf{z}'\boldsymbol{V}(\boldsymbol{\theta})\mathbf{z} + \sum_{i=1}^{n} \ln p(\mathbf{y}_i|\mathbf{z}_i,\boldsymbol{\theta})\right).$$

Rue, et al (2004) and Rue, et al (2009) showed that this type of posterior distribution can be well approximated by a Gaussian distribution that matches the mode and the curvature at the mode. The resulting approximation is known as the Laplace approximation and can be expressed as,

$$p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}) \propto \exp\left(-\frac{1}{2}\mathbf{z}'(V(\boldsymbol{\theta}) + diag(\mathbf{c}))\mathbf{z}\right),$$

where $\mathbf{c}$ comes from the second order term in the Taylor expansion of $\sum_{i=1}^{n} \ln p(\mathbf{y}_i|\mathbf{z}_i)$ at the mode of $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$. The Laplace approximation may be employed to compute $\mathcal{H}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}})$. After $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is obtained, it is easy to obtain $D(\bar{\boldsymbol{\theta}})$. It is important to point out that the numerical evaluation of $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is needed only once, i.e., at the posterior mean.

To compute $P_D^*$, we have to calculate the second derivative of the observed-data likelihood function. The following two methods can be used. First, if the $\mathcal{Q}$ function is available in closed-form, we can use the following formula given in Oakes (1999),

$$\frac{\partial^2 \mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \left\{\frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial^2 \mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{*\prime}}\right\}_{\boldsymbol{\theta}^*=\boldsymbol{\theta}}. \tag{20}$$

Second, if the $\mathcal{Q}$ function does not have an analytic form, Louis (1982) showed that that

$$\begin{aligned}\frac{\partial^2 \mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\} + Var_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{S(\mathbf{x}|\boldsymbol{\theta})\right\} \qquad (21)\\ &= E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{x}|\boldsymbol{\theta})S(\mathbf{x}|\boldsymbol{\theta})'\right\} - E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{x}|\boldsymbol{\theta})\}E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{x}|\boldsymbol{\theta})\}',\end{aligned}$$

where $S(\mathbf{x}|\boldsymbol{\theta}) = \partial\mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ and all the expectations are taken with respect to the conditional distribution of $\mathbf{z}$ given $\mathbf{y}$ and $\boldsymbol{\theta}$. Hence, we can use the following formula to calculate the second derivative of the observed-data likelihood function,

$$\begin{aligned}E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\left\{\frac{\partial^2 \mathcal{L}_c(\mathbf{x}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{x}|\boldsymbol{\theta})S(\mathbf{x}|\boldsymbol{\theta})'\right\} &\approx \frac{1}{M}\sum_{m=1}^{M}\left\{\frac{\partial^2 \mathcal{L}_c(\mathbf{y},\mathbf{z}^{(m)}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + S(\mathbf{y},\mathbf{z}^{(m)}|\boldsymbol{\theta})S(\mathbf{y},\mathbf{z}^{(m)}|\boldsymbol{\theta})'\right\},\\ E_{\mathbf{z}|\mathbf{y},\boldsymbol{\theta}}\{S(\mathbf{x}|\boldsymbol{\theta})\} &\approx \frac{1}{M}\sum_{m=1}^{M} S(\mathbf{y},\mathbf{z}^{(m)}|\boldsymbol{\theta}), \qquad (22)\end{aligned}$$

where $\{\mathbf{z}^{(m)}, m = 1, 2, \cdots, M\}$ are random observations drawn from the posterior distribution $p(\boldsymbol{z}|\mathbf{y}, \boldsymbol{\theta})$.

# 5 Examples

We now illustrate the proposed method in two applications. In the first example, while $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is not available in closed-form, the Kalman filter provides a recursive algorithm to evaluate it. Hence, $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta})$ and $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta})$ can be calculated in the same manner, facilitating the computation of RDIC. In the second example, $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is not available in closed-form and the Kalman filter cannot be applied. To compute RDIC, we use the Laplace approximation and the technique suggested in (22).

## 5.1 Comparing high dimensional dynamic factor models

For many countries, there exists a rich array of macroeconomic time series and financial time series. To reduce the dimensionality and to extract the information from the large number of time series, factor analysis has been widely used in the empirical macroeconomic literature and in the empirical finance literature. For example, by extending the static factor models previously developed for cross-sectional data, Geweke (1977) proposed the dynamic factor model for time series data. Many empirical studies, such as Sargent and Sims (1977), Giannone, et al (2004), have reported evidence that a large fraction of the variance of many macroeconomic series can be explained by a small number of dynamic factors. Stock and Watson (1999) and Stock and Watson (2002) showed that dynamic factors extracted from a large number of predictors lead to improvement in predicting macroeconomic variables. Not surprisingly, high dimensional dynamic factor models have become a popular tool under a data rich environment for macroeconomists and policy makers. An excellent review on the dynamic factor models is given by Stock and Watson (2011).

Following Bernanke, et al (2005) (BBE hereafter), the present paper considers the following fundamental dynamic factor model:

$$\begin{aligned} Y_t &= F_t L' + \varepsilon_t', \\ F_t &= F_{t-1}\Phi' + \eta_t, \end{aligned}$$

where $Y_t$ is a $1 \times N$ vector of time series variables, $F_t$ a $1 \times K$ vector of unobserved latent factors which contains the information extracted from all the $N$ time series variables, $L$ an $N \times K$ factor loading matrix, $\Phi$ the $K \times K$ autoregressive parameter matrix of unobserved latent factors. It is assumed that $\varepsilon_t \sim N(0, \Sigma)$ and $\eta_t \sim N(0, Q)$. For the purpose of identification, $\Sigma$ is assume to be diagonal and $\varepsilon_t$ and $\eta_t$ are assumed to be independent with each other. Following BBE (2005), we set the first $K \times K$ block in the loading matrix $L$ to be the identity matrix.

In this dynamic factor model, the observed variable $Y_t$ consists of a balanced panel of 120 US monthly macroeconomic time series. These series were transformed to induce stationarity by BBE (2005). The description of the series and the transformation is provided in BBE (2005). The sample period is from January 1959 to August 2001. Because the data are of high dimension, the analysis of the dynamic factor models via a frequentist method is difficult; see the discussion in Stock and Watson (2011). In the literature, the Bayesian inference via the MCMC techniques has been popular for analyzing the dynamic factor models; see Otrok and Whiteman (1998), Kose, et al (2003, 2008), BBE (2005).

Following BBE (2005), we specify the following prior distributions:

$$\begin{aligned} \Sigma_{ii} &\sim Inverse-\Gamma\left(3, 0.001\right), L_i \sim N\left(0, \Sigma_{ii} M_0^{-1}\right), \\ vec\left(\Phi\right)|Q &\sim N\left(0, Q \otimes \Omega_0\right), Q \sim Inverse-\Gamma\left(Q_0, K+2\right), \end{aligned}$$

where $M_0$ is a $K \times K$ identity matrix, $L_i$ the $i$th ($i > K$) column of $L$. The diagonal elements of $Q_0$ are set to be the residual variances of the corresponding AR(1) model, $\left\{\hat{\sigma}_i^2\right\}$. The diagonal elements of $\Omega_0$ are constructed so that the prior variance of the parameter on the $j$th variable in the $i$th equation is $\hat{\sigma}_i^2/\hat{\sigma}_j^2$.

In this example, we aim to determine the number of factors in the dynamic factor models using model selection criteria. In BBE (2005) model comparison is achieved by graphic methods. Our approach can be regarded as a formal statistical alternative to graphic methods. It is well documented that the determination of number of factors in dynamic factor models is important; see Stock and Watson (1999). As in the previous example, we use $\text{DIC}_7$ and RDIC to compare models with different numbers of factors, namely $K = 1$, 2 and 3, which are denoted by $M_1$, $M_2$, $M_3$ respectively. 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.

Following a suggestion of a referee, we also compare alternative models using the marginal likelihood approach. Unfortunately, the prior distributions of $\Phi$ and $Q$ of BBE (2005) depend on the latent variables which lead to implicit joint prior distributions of $L, R, \Phi$ and $Q$. Consequently, it is difficult to calculate the joint prior density of $L, R, \Phi$ and $Q$. To avoid the evaluation of the joint prior density, we calculate the marginal likelihood by the harmonic mean method (Newton and Raftery, 1994), which only needs to calculate the reciprocal of the likelihood for each posterior draw of parameters.

Based on the 20,000 samples, we compute $\text{DIC}_7$, RDIC, and the marginal likelihood for all three models. Equation (18) is used to approximate the observed-data likelihood at the posterior mean**.** Table 1 reports the simple count of the number of parameters (including the latent variables), $\text{DIC}_7$, the $P_D$ component of $\text{DIC}_7$, (i.e. when the data augmentation technique is used), the simple count of the number of parameters (excluding the latent variables),

Table 1: Model selection results for dynamic factor models

| Model | $M_1$ | $M_2$ | $M_3$ |
|---|---|---|---|
| Number of Parameters | 752 | 1385 | 2019 |
| $P_D$ | 354 | 971 | 1404 |
| $\text{DIC}_7$ | -23288 | -37851 | -44568 |
| Number of Parameters | 241 | 363 | 486 |
| $P_D^*$ | 88 | 203 | 316 |
| $D\left(\overline{\theta}\right)$ | -22594 | -35248 | -41015 |
| RDIC | -22418 | -34842 | -40383 |
| Log MargLik | 10733 | 16978 | 19842 |

RDIC, the $P_D^*$ component and the $D\left(\overline{\boldsymbol{\theta}}\right)$ component of RDIC (i.e. when the data augmentation technique is not used), and the marginal likelihood. Several conclusions may be drawn from Table 1. First, $\text{DIC}_7$, RDIC and the marginal likelihood all suggest that $M_3$ is the best model, followed by Model 2 and then by Model 1. Model 3 has a higher effective number of parameters than the other two models. However, the gain in the fit to data is greater. The conclusion is that at least 3 factors are needed to describe the joint movement of the 120 macroeconomic time series. Second, since very informative priors have been used, neither $P_D$ nor $P_D^*$ is close to the actual number of parameters. While it is cheap to compute RDIC, it is much harder to compute $\text{DIC}_1$. This is because the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form and the Kalman filter is used to numerically calculate $p(\mathbf{y}|\boldsymbol{\theta})$ which involves the computation of $\frac{1}{J}\sum_{j=1}^{J}\ln p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$, for $J = 20,000$. We have to run the Kalman filter 20,000 times, which takes more than 4 hours to compute in Matlab.[4] In sharp contrast, it only took less than 80 seconds to compute RDIC. Obviously, the discrepancy in CPU time increases with $J$.

### 5.2 Comparing stochastic volatility models

Stochastic volatility (SV) models have been found very useful for pricing derivative securities. In the discrete time log-normal SV models, the logarithmic volatility is the state variable which is often assumed to follow an AR(1) model. The basic log-normal SV model is of the form:

$$y_t = \alpha + \exp(h_t/2)u_t,\ u_t \sim N(0,1),$$
$$h_t = \mu + \phi(h_{t-1} - \mu) + v_t,\ v_t \sim N(0,\tau^2),$$

where $t = 1, 2, \cdots, n$, $y_t$ is the continuously compounded return, $h_t$ the unobserved log-volatility, $h_0 = \mu$, $u_t$ and $v_t$ are independent for all $t$. In this paper, we denote this model by

[4]Numerically more efficient algorithms, such as the one proposed by Chan and Jeliazkov (2009) may be used to evaluate $\ln p(\mathbf{y}|\boldsymbol{\theta}^{(j)})$.

$M_1$.

To carry out Bayesian analysis of $M_1$, following Meyer and Yu (2000), the prior distributions are specified as follows:

$$\begin{aligned} \alpha &\sim N(0, 100), \quad \mu \sim N(0, 100), \\ \phi &\sim Beta(1, 1), \quad 1/\tau^2 \sim \Gamma(0.001, 0.001). \end{aligned}$$

An important and well documented empirical feature in many financial time series is the leverage effect (Black, 1976). Following Yu (2005), we define the leverage effect SV model as:

$$\begin{aligned} y_t &= \alpha + \exp(h_t/2)\, u_t, \ u_t \sim N(0, 1) \\ h_{t+1} &= \mu + \phi(h_t - \mu) + v_{t+1}, \ v_{t+1} \sim N\left(0, \tau^2\right) \end{aligned}$$

with

$$\begin{pmatrix} u_t \\ v_{t+1} \end{pmatrix} \overset{i.i.d}{\sim} N\left\{ \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right\}$$

and $h_0 = \mu$. In this model, $\rho$ captures the leverage effect if $\rho < 0$. In this case, there is a negative relationship between the expected future volatility and the current return. We denote this model as $M_2$ and specify the prior distribution of $\rho$ as:

$$\rho \sim \text{Unif}(-1, 1).$$

Our goal here is to compare the two models using $\text{DIC}_7$, RDIC and the Bayes factor $BF_{21}$. In all cases, $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form. Since the models are of a nonlinear non-Gaussian form, the Kalman filter cannot be applied and $\text{DIC}_1$ is infeasible to compute. Specifically, for nested models, the Bayes factor can be calculated using the Savage Dickey density ratio (Verdinelli and Wasserman, 1995).

The dataset consists of 945 daily mean-corrected returns on Pound/Dollar exchange rates, covering the period between 01/10/81 and 28/06/85. For MCMC, after a burn-in period of 10,000 iterations, we save every 20th value for the next 100,000 iterations to get 5,000 effective draws. The same dataset was used in Kim, Shephard and Chib (1998) and Meyer and Yu (2000). The posterior mean and standard error of parameters in the two competing model are reported in Table 2. Note that the in $M_2$, the posterior mean of $\rho$ is very close to zero, relative to its posterior standard error.

Table 3 reports $\text{DIC}_7$, RDIC, $P_D$, $P_D^*$ and the Savage-Dickey density ratio for the two models. Since the $\mathcal{Q}$ function does not have a closed-form expression, we employ Equations (21) and (22) to compute the second order derivative of the observed-data likelihood. To compute RDIC, we use the Laplace approximation of Rue, Martino and Chopin (2009). Equation (18) is used to approximate the observed-data likelihood at the posterior mean. In particular, we run the Gibbs sampling twice, one for the parameters and the latent variables, another

Table 2: Posterior mean and standard error of parameters in $M_1$ and $M_2$

| | $M_1$ | | $M_2$ | |
|---|---|---|---|---|
| Parameter | Mean | SE | Mean | SE |
| $\mu$ | -0.6733 | 0.3282 | -0.6485 | 0.3377 |
| $\phi$ | 0.9733 | 0.0127 | 0.9802 | 0.0138 |
| $\rho$ | - | - | -0.0575 | 0.1570 |
| $\tau$ | 0.1698 | 0.0378 | 0.1661 | 0.0391 |

Table 3: Model selection results for $M_1$ and $M_2$

| Model | $M_1$ | $M_2$ |
|---|---|---|
| $P_D$ | 53.60 | 31.33 |
| $D(\bar{\boldsymbol{\theta}})$ | 1695.40 | 1693.36 |
| $\text{DIC}_7$ | 1802.52 | 1756.21 |
| $P_D^*$ | 2.32 | 3.24 |
| $D(\bar{\boldsymbol{\theta}})$ | 1837.81 | 1837.78 |
| RDIC | 1842.50 | 1844.30 |
| $BF_{21}$ | 0.2174 | |

for the latent variables given the parameters at the posterior mean obtained from the earlier Gibbs sampler.

The following findings can be obtained from Table 3. First and foremost, $\text{DIC}_7$ and RDIC suggest different rankings of the competing models. In particular, by dropping the value by 43.3, $\text{DIC}_7$ suggests that $M_2$ is better that $M_1$. According to $\text{DIC}_7$, $M_1$ and $M_2$ perform nearly the same judged by $D(\bar{\boldsymbol{\theta}})$. However, $M_2$ reduces $P_D$ by 22.3 over $M_1$. This reduction of the model complexity is the reason why $\text{DIC}_7$ prefers $M_2$. This result is surprising as the posterior mean of the leverage effect is nearly zero as reported in Table 2. On the other hand, RDIC suggests that $M_1$ is slightly better that $M_2$ although the difference is not worth to mention. In RDIC, $P_D^*$ is 2.32 in $M_1$ and 3.24 in $M_2$. These values are very close to the actual numbers of parameters in the two models. Given that $M_2$ has one extra parameter, this difference is reasonable. Moreover, $M_1$ and $M_2$ perform nearly the same judged by $D(\bar{\boldsymbol{\theta}})$. These two observations explain why $M_1$ is slightly better that $M_2$. Third, Bayes factor suggest that $M_1$ is the better model, consistent with the ranking of RDIC. This empirical example clearly demonstrates that RDIC is a more reliable model selection criterion that $\text{DIC}_7$.

# 6 Conclusion

This paper provides a rigorous decision-theoretic justification of DIC when there is no latent variable in candidate models. Although latent variable models can be conveniently estimated in the Bayesian framework via MCMC if the data augmentation technique is used, we argue that data augmentation cannot be used in connection to DIC. This is because the justification of DIC rests on the validity of the standard Bayesian asymptotic theory. With data augmentation, the number of parameters increases with the number of observations, invalidating the standard Bayesian large sample theory. In addition, the use of the data augmentation makes DIC very sensitive to transformations and distributional representations.

While in principle one can use the standard DIC (i.e. $\text{DIC}_1$) without resorting to the data augmentation technique, in practice this standard DIC is very difficult to use because the observed-data likelihood is not available in closed-form for many latent variable models and the standard $\text{DIC}_1$ has to numerically evaluate the observed-data likelihood at each MCMC iteration. These two observations make the implementation of $\text{DIC}_1$ practically non-operational for latent variable models.

We introduces a robust deviance information criteria (RDIC) for comparing models with latent variables. RDIC is defined without augmenting the parameter space and hence can be justified by the standard Bayesian asymptotic theory. We then show how the EM algorithm can facilitate the computation of RDIC in different contexts. Since the latent variables are not treated as parameters in our approach, RDIC is robust to nonlinear transformations of the latent variables and distributional representations of the model specification. Asymptotic justification, computational tractability and robustness to transformation and specification are the three main advantages of the proposed approach. These advantages are illustrated using two popular models in economics and finance.

Both $\text{DIC}_1$ and RDIC require that the DGP be nested by the candidate model and that the standard ML theory holds true. How to develop a good information criterion for comparing latent variable models, with the possibility that the candidate model is misspecified, will be pursued in future research. Also, the topic on comparing models for which the standard ML theory fails will be pursued in future studies.

# Appendix

## Proof of Lemma 3.1

Under Assumptions 1-5, for any $\epsilon > 0$, let $n > \max\{n^*, n^{**}\}$ and $\delta > 0$, for any $\boldsymbol{\theta} \in H\left(\hat{\boldsymbol{\theta}}_m, \delta\right) = \left\{\boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m|| \leq \delta\right\}$, we have

$$\begin{aligned}\ln p(\boldsymbol{\theta}|\mathbf{y}) &= \ln p(\hat{\boldsymbol{\theta}}_m|\mathbf{y}) + L_n^{(1)}(\hat{\boldsymbol{\theta}}_m)'(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) \\ &= \ln p(\hat{\boldsymbol{\theta}}_m|\mathbf{y}) + \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m),\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}$ lies on the segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_m$. It follows that

$$p(\boldsymbol{\theta}|\mathbf{y}) = p(\hat{\boldsymbol{\theta}}_m|\mathbf{y}) \exp\left[\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)\right].$$

Let $\boldsymbol{\omega} = \sqrt{n}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)$, $J(\boldsymbol{\theta}) = -\frac{1}{n} L_n^{(2)}(\boldsymbol{\theta})$. For given $\epsilon$ and $\delta$ such that $\Omega = \{\boldsymbol{\omega} : ||\boldsymbol{\omega}|| < \sqrt{n}\delta\}$, we have $\boldsymbol{\theta} \in H(\hat{\boldsymbol{\theta}}_m, \delta)$. It can be shown that

$$p(\boldsymbol{\omega}|\mathbf{y}) \propto \exp\left[\frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\tilde{\boldsymbol{\theta}})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)\right] = \exp\left\{-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right\}.$$

Let $c_n^* = \int_\Omega \exp[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}]\mathrm{d}\boldsymbol{\omega}$, $c_n = \int_\Omega \exp[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}]\mathrm{d}\boldsymbol{\omega}$, we have

$$\begin{aligned}
& P_n := \int_\Omega \left| p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \int_\Omega \left| \frac{1}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \frac{1}{c_n}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \frac{1}{c_n}\int_\Omega \left| \frac{c_n}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \frac{1}{c_n}\int_\Omega \left| \frac{c_n - c_n^*}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] + \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{1}{c_n}\left\{\int_\Omega \left|\frac{c_n - c_n^*}{c_n^*}\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega} + \int_\Omega \left|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega}\right\} \\
\leq & \frac{|c_n - c_n^*|}{c_n} + \frac{1}{c_n}\int_\Omega \left|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{2}{c_n}\int_\Omega \left|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{2}{c_n}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}}) - J(\hat{\boldsymbol{\theta}}_m)\right]\boldsymbol{\omega}\right\} - 1\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

When $\Omega = \{\boldsymbol{\omega} : ||\boldsymbol{\omega}|| < \sqrt{n}\delta\}$, we have $\boldsymbol{\theta} \in H(\hat{\boldsymbol{\theta}}_m, \delta)$ and $-A(\epsilon) \leq [J(\tilde{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}}_m) - I_P] \leq A(\epsilon)$.

By the Hölder inequality, we have

$$\begin{aligned}
&\lim_{n\to\infty} Q_n := \lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})-J(\hat{\boldsymbol{\theta}}_m)\right]\boldsymbol{\omega}\right\}-1\right|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}\\
=\;& \lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})-J(\hat{\boldsymbol{\theta}}_m)\right]\boldsymbol{\omega}\right\}-1\right|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}\\
=\;& \lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}}_m)-I_P\right]J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right\}-1\right|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}\\
\le\;& \left\{\lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}}_m)-I_P\right]J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right\}-1\right|^2\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}\right\}^{1/2}\\
=\;& (D_1-2D_2+D_3)^{1/2},
\end{aligned}$$

where

$$\begin{aligned}
D_1 &= \lim_{n\to\infty}\int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega},\\
D_2 &= \lim_{n\to\infty}\int_\Omega \exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}}_m)-I_P\right]J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right\}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}\\
&= \lim_{n\to\infty}\int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega},\\
D_3 &= \lim_{n\to\infty}\int_\Omega \exp\left\{-\boldsymbol{\omega}'\left[J(\tilde{\boldsymbol{\theta}})J^{-1}(\hat{\boldsymbol{\theta}}_m)-I_P\right]J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right\}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}\right]\mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

It can be shown that $D_1 = (2\pi)^{P/2}|J(\hat{\boldsymbol{\theta}}_m)|^{-1/2}$. Following the proof of Lemma 2.1 and Theorem 2.1 of Chen (1985), we have $D_2^- \le D_2 \le D_2^+, D_3^- \le D_3 \le D_3^+$ and

$$\begin{aligned}
D_2^+ &= \left|J(\hat{\boldsymbol{\theta}}_m)\right|^{-1/2}|I_P - A(\epsilon)|^{-1/2}\int_{||Z||<s_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z},\\
D_2^- &= \left|J(\hat{\boldsymbol{\theta}}_m)\right|^{-1/2}|I_P + A(\epsilon)|^{-1/2}\int_{||Z||<t_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z},\\
D_3^+ &= \left|J(\hat{\boldsymbol{\theta}}_m)\right|^{-1/2}|I_P - 2A(\epsilon)|^{-1/2}\int_{||Z||<s'_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z},\\
D_3^- &= \left|J(\hat{\boldsymbol{\theta}}_m)\right|^{-1/2}|I_P + 2A(\epsilon)|^{-1/2}\int_{||Z||<t'_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z},
\end{aligned}$$

where $s_n = \delta(1-e^*(\epsilon))^{1/2}/\sigma_n^*$, $t_n = \delta(1+e^*(\epsilon))^{1/2}/\sigma_n$, $s'_n = \delta(1-2e^*(\epsilon))^{1/2}/\sigma_n^*$ and $t'_n = \delta(1+2e^*(\epsilon))^{1/2}/\sigma_n$, $\sigma_n^2$ and $\sigma_n^{*2}$ is the largest and smallest eigenvalue of $\{J(\hat{\boldsymbol{\theta}}_m)\}^{-1}$, $e(\epsilon)$ and $e^*(\epsilon)$ is the largest and smallest eigenvalue of $A(\epsilon)$. Under the regularity conditions, when $n\to\infty$, $s_n\to\infty$, $t_n\to\infty$, $s'_n\to\infty$, $t'_n\to\infty$, then if $\epsilon\to 0$, we get

$$\begin{aligned}
&\lim_{n\to\infty}|I_P \pm A(\epsilon)| = 1,\ \lim_{n\to\infty}|I_P \pm 2A(\epsilon)| = 1,\\
&\lim_{n\to\infty}\int_{||Z||<s_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z} = (2\pi)^{P/2},\\
&\lim_{n\to\infty}\int_{||Z||<t_n}\exp\left[-\frac{1}{2}\boldsymbol{Z}'\boldsymbol{Z}\right]\mathrm{d}\boldsymbol{Z} = (2\pi)^{P/2}.
\end{aligned}$$

Then, we can show that $D_1 = D_2 = D_3 = (2\pi)^{P/2}|J(\hat{\boldsymbol{\theta}}_m)|^{-1/2}$ which implies that $\lim_{n\to\infty} Q_n = 0$ and $\lim_{n\to\infty} P_n = 0$.

For $i, j = 1, 2, \cdots, P$, it can be shown that

$$\begin{aligned}
&\lim_{n\to\infty} \left| \int_\Omega \omega_i \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} \right| \\
\leq\ & \lim_{n\to\infty} \int_\Omega \left| \omega_i \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \right\} \right| \mathrm{d}\boldsymbol{\omega} \\
\leq\ & \lim_{n\to\infty} \frac{|c_n - c_n^*|}{c_n} \int |\omega_i| p(\boldsymbol{\omega}|\mathbf{y}) \mathrm{d}\boldsymbol{\omega} \\
+\ & \lim_{n\to\infty} \frac{1}{c_n} \int_\Omega |\omega_i| \left| \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[ J(\tilde{\boldsymbol{\theta}}) - J(\hat{\boldsymbol{\theta}}_m) \right] \boldsymbol{\omega} \right\} - 1 \right| \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}
\end{aligned}$$

By Assumption 5, it can be shown that

$$\frac{|c_n - c_n^*|}{c_n} \int |\omega_i| p(\boldsymbol{\omega}|\mathbf{y}) \mathrm{d}\boldsymbol{\omega} \longrightarrow 0$$

By using the Hölder's inequality, it also can be shown that

$$\begin{aligned}
&\int_\Omega |\omega_i| \left| \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[ J(\tilde{\boldsymbol{\theta}}) - J(\hat{\boldsymbol{\theta}}_m) \right] \boldsymbol{\omega} \right\} - 1 \right| \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
\leq\ & \left\{ \int_\Omega \left| \exp\left\{ -\frac{\boldsymbol{\omega}' \left[ J(\tilde{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}_m) - I_P \right] J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}}{2} \right\} - 1 \right|^2 \exp\left[ -\frac{\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}}{2} \right] \mathrm{d}\boldsymbol{\omega} \right\}^{\frac{1}{2}} \\
&\times \left\{ \int_\Omega \omega_i^2 \exp\left[ -\frac{\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega}}{2} \right] \mathrm{d}\boldsymbol{\omega} \right\}^{\frac{1}{2}} \\
=\ & \sqrt{E(\omega_i^2)} (ED_1 - 2ED_2 + ED_3)^{1/2} \longrightarrow 0
\end{aligned}$$

where

$$\begin{aligned}
ED_1 &= \int_\Omega \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}, \\
ED_2 &= \int_\Omega \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[ J(\tilde{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}_m) - I_P \right] J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right\} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
&= \int_\Omega \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\tilde{\boldsymbol{\theta}})\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}, \\
ED_3 &= \int_\Omega \omega_i^2 \exp\left\{ -\boldsymbol{\omega}' \left[ J(\tilde{\boldsymbol{\theta}}) J^{-1}(\hat{\boldsymbol{\theta}}_m) - I_P \right] J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right\} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
&= \int_\Omega \omega_i^2 \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[ 2J(\tilde{\boldsymbol{\theta}}) - J(\hat{\boldsymbol{\theta}}_m) \right] J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right\} \mathrm{d}\boldsymbol{\omega}
\end{aligned}$$

and $ED_1 - 2ED_2 + ED_3 \longrightarrow 0$

Hence, we have

$$\begin{aligned}
&\left| \int_\Omega \omega_i \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} \right| \\
\leq\ & \int_\Omega |\omega_i| \left| p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2}\boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m)\boldsymbol{\omega} \right] \right| \mathrm{d}\boldsymbol{\omega} \longrightarrow 0,
\end{aligned}$$

Similarly, we also can show that

$$\begin{aligned} & \left| \int_{\Omega} \omega_i \omega_j \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2} \boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m) \boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} \right| \\ \leq & \int_{\Omega} |\omega_i \omega_j| \left| p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[ -\frac{1}{2} \boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m) \boldsymbol{\omega} \right] \right| \mathrm{d}\boldsymbol{\omega} \longrightarrow 0. \end{aligned}$$

Note that

$$\begin{aligned} \lim_{n\to\infty} \int_{\Omega} \omega_i \left\{ \frac{1}{c_n} \exp\left[ -\frac{1}{2} \boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m) \boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} = 0, \\ \lim_{n\to\infty} \int_{\Omega} \omega_i \omega_j \left\{ \frac{1}{c_n} \exp\left[ -\frac{1}{2} \boldsymbol{\omega}' J(\hat{\boldsymbol{\theta}}_m) \boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} = J^{-1}_{ij}(\hat{\boldsymbol{\theta}}_m), \end{aligned}$$

where $J^{-1}_{ij}(\hat{\boldsymbol{\theta}}_m)$ is the $(i,j)^{th}$ element of $J^{-1}(\hat{\boldsymbol{\theta}}_m)$. Hence, given the observed data $\mathbf{y}$, $E(\boldsymbol{\omega}|\mathbf{y}) = 0 + o(1)$ and $E(\boldsymbol{\omega}\boldsymbol{\omega}'|\mathbf{y}) = J^{-1}(\hat{\boldsymbol{\theta}}_m) + o(1)$ which imply that

$$E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)|\mathbf{y}] = o_p(n^{-1/2}), E[(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)'|\mathbf{y}] = -L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1}).$$

## Proof of Theorem 3.1

Under Assumption 1, we know that $-L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) = O_p(n)$. Under Assumption 3, for any $\epsilon > 0$, let $n > \max\{n^*, n^{**}\}$, there exists an integer $\delta > 0$, for any $\boldsymbol{\theta} \in H\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}, \delta\right)$, we have

$$-A(\epsilon) \leq L_n^{(2)}(\boldsymbol{\theta}) L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) - \mathbf{I}_P \leq A(\epsilon).$$

As $n \to \infty$, when $\epsilon \to 0$ and $A(\epsilon) \to 0$, we get $L_n^{(2)}(\boldsymbol{\theta}) L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) - \mathbf{I}_P = o_p(1)$. Furthermore, for any $\boldsymbol{\theta} \in H\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}, \delta\right)$, we get

$$\begin{aligned} & -L_n^{(2)}(\boldsymbol{\theta}) + L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) = \left[ L_n^{(2)}(\boldsymbol{\theta}) L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) - \mathbf{I}_P \right] \left[ -L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) \right] = o_p(1) O_p(n) = o_p(n), \\ & L_n^{(2)}(\boldsymbol{\theta}) = L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) + o_p(n) = O_p(n). \end{aligned}$$

Since, $p(\boldsymbol{\theta}) = O_p(1)$, hence, we can get that

$$\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = L_n^{(2)}(\boldsymbol{\theta}) - \frac{\partial^2 \ln p(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = O_p(n) - O_P(1) = O_p(n).$$

For any $\epsilon > 0$, let $n > \max\{n^*, n^{**}\}$ and $\delta > 0$, for any $\boldsymbol{\theta} \in H\left(\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}, \delta\right) = \left\{ \boldsymbol{\theta} : ||\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}|| \leq \delta \right\}$, using the Taylor expansion, we can get

$$p(\mathbf{y}_{rep}|\boldsymbol{\theta}) = p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + \frac{\partial p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta}} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + \frac{1}{2} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})' \frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}),$$

where $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$ lies on the segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$. Hence, we can show that

$$\begin{aligned}
& p(\mathbf{y}_{rep}|\mathbf{y}) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
= \; & p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + \frac{\partial p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta}} \int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
& + \frac{1}{2} \int \left[ (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})' \frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
= \; & p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta}} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \\
& + \frac{1}{2} p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \int \left[ (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})' \frac{1}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})} \frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
= \; & p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})(1 + c_1 + c_2),
\end{aligned}$$

where

$$\begin{aligned}
c_1 &= \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta}} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}), \\
c_2 &= \frac{1}{2} \int \left[ (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})' \frac{1}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})} \frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) \right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}.
\end{aligned}$$

Next we will show that $c_1 = o_p(1)$ and $c_2 = o_p(1)$. When the model is correctly specified, the quasi-true value $\boldsymbol{\theta}_t$ is the true value denoted by $\boldsymbol{\theta}_0$. Since $\mathbf{y}_{rep}$ is the replicate data, under Assumption 6, it can be shown that

$$\frac{1}{\sqrt{n}} \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}} \sim N\left[0, \mathbf{J}\left(\boldsymbol{\theta}_0\right)\right],$$

Thus,

$$\frac{1}{\sqrt{n}} \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta_0})}{\partial \boldsymbol{\theta}} = O_p(1)$$

Since $\hat{\boldsymbol{\theta}}_m$ is the consistent estimator of $\boldsymbol{\theta}_0$, using continuous mapping theorem, we get

$$\frac{1}{\sqrt{n}} \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_m)}{\partial \boldsymbol{\theta}} = \frac{1}{\sqrt{n}} \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}} + o_p(1) = O_p(1).$$

Using Lemma 3.1, we get

$$c_1 = \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial \boldsymbol{\theta}} (\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) = O_p(n^{1/2}) o_p(n^{-1/2}) = o_p(1).$$

According to the information matrix equality for the correctly specified model and Assumption 6, we have $\mathbf{H}(\boldsymbol{\theta}_0) = \mathbf{J}(\boldsymbol{\theta}_0)$. Using the standard Bayesian and ML large sample theory, we can have $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} = \hat{\boldsymbol{\theta}}_{\boldsymbol{m}} + o_p(1)$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}} = \boldsymbol{\theta}_0 + o_p(1)$ so that $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$ is also a consistent estimator of $\boldsymbol{\theta}_0$. Hence, using the continuous mapping theorem, we can get

$$\mathbf{H}(\tilde{\boldsymbol{\theta}}_m) = \mathbf{H}(\boldsymbol{\theta}_0) + o_p(1), \; \mathbf{J}(\tilde{\boldsymbol{\theta}}_m) = \mathbf{J}(\boldsymbol{\theta}_0) + o_p(1).$$

Furthermore, we can show

$$-\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{H}(\tilde{\boldsymbol{\theta}}_m) + o_p(1) = \mathbf{H}(\boldsymbol{\theta}_0) + o_p(1),$$
$$\frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}'} = \mathbf{J}(\tilde{\boldsymbol{\theta}}_m) + o_p(1) = \mathbf{J}(\boldsymbol{\theta}_0) + o_p(1).$$

Note that

$$\frac{1}{p(\mathbf{y}_{rep}|\boldsymbol{\theta})}\frac{\partial^2 p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\boldsymbol{\partial\theta}\boldsymbol{\partial\theta}'} = \frac{\partial^2 \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}$$

Hence, we can get

$$\begin{aligned}&\frac{1}{n}\frac{1}{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}\frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\boldsymbol{\partial\theta}\boldsymbol{\partial\theta}'} = \frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}'}\\ =\ & -\mathbf{H}(\boldsymbol{\theta}_0) + \mathbf{J}(\boldsymbol{\theta}_0) + o_p(1) = o_p(1).\end{aligned}$$

There exists some null sequence $k_n$, let

$$a_n = \sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_t||\le k_n} || - \frac{1}{n}\left[\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] - \mathbf{H}(\boldsymbol{\theta}_0)||$$
$$b_n = \sup_{||\boldsymbol{\theta}-\boldsymbol{\theta}_t||\le k_n} ||||\frac{1}{n}\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}' - \mathbf{J}(\boldsymbol{\theta}_0)||$$

Furthermore, since $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$ is also a consistent estimator of $\boldsymbol{\theta}_0$, by assumption 6, we can get that

$$\left|\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - (-\mathbf{H}(\boldsymbol{\theta}_0))\right| \le a_n\mathbf{1}_P \xrightarrow{p} 0$$
$$\left|\frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}'} - \mathbf{J}(\boldsymbol{\theta}_0)\right| \le b_n\mathbf{1}_P \xrightarrow{p} 0$$

where $\mathbf{1}_P$ is a $P \times P$ matrix with one as every component. Hence, we can get that

$$\begin{aligned}&\left|\frac{1}{n}\frac{1}{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}\frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\boldsymbol{\partial\theta}\boldsymbol{\partial\theta}'} - 0\times\mathbf{1}_P\right|\\ &\left|\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}'} + \mathbf{H}(\boldsymbol{\theta}_0) - \mathbf{J}(\boldsymbol{\theta}_0)\right|\\ \le\ &\left|\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - (-\mathbf{H}(\boldsymbol{\theta}_0))\right| + \left|\frac{1}{n}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}\frac{\partial \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}'} - \mathbf{J}(\boldsymbol{\theta}_0)\right|\\ \le\ &(a_n + b_n)\mathbf{1}_P \xrightarrow{p} 0,\end{aligned}$$

Under the Bayesian large sample theory, we know that $\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}} = O_p(n^{-1/2})$ so that $\tilde{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_m = O_p(n^{-1/2})$. Using the Taylor expansion, we have

$$\begin{aligned}&\ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\\ =\ & \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + \frac{1}{2}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})'\frac{\partial^2 \ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m1}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\\ =\ & O_p(n^{1/2})O_p(n^{-1/2}) + O_p(n^{-1/2})O_p(n)O_p(n^{-1/2}) = O_p(1),\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m1}}$ lies on the segment between $\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}$ and $\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}$. Then, we get

$$\frac{p(\mathbf{y}|\tilde{\boldsymbol{\theta}}_m)}{p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)} = \exp\left[\ln p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}}) - \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\right] = \exp(O_p(1)) = O_p(1),$$

Hence, we can have

$$\begin{aligned}
|c_2| &= \left|\frac{1}{2}\int\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})'\frac{1}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}\frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\right]p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right| \\
&= \left|\frac{1}{2}\int\left[n(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})'\frac{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}\left[\frac{1}{n}\frac{1}{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}\frac{\partial^2 p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right](\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\right]p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right| \\
&\leq \left|\frac{1}{2}\int\left[n(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})'\frac{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}\left[(a_n+b_n)\mathbf{1}_P\right](\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\right]p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right| \\
&= (a_n+b_n)\left|\frac{1}{2}\int\left[\sqrt{n}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})'\frac{p(\mathbf{y}_{rep}|\tilde{\boldsymbol{\theta}}_{\boldsymbol{m}})}{p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})}\mathbf{1}_P\sqrt{n}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_{\boldsymbol{m}})\right]p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}\right| \xrightarrow{p} 0
\end{aligned}$$

Let $\hat{\boldsymbol{\theta}}_{ML}$ be the ML estimator of $\boldsymbol{\theta}$. Using the Taylor expansion, $p(\boldsymbol{\theta}) = O_p(1)$, we can get

$$0 = \frac{\partial \ln p(\mathbf{y},\hat{\boldsymbol{\theta}}_m)}{\partial\boldsymbol{\theta}} = \frac{\partial \ln p(\mathbf{y},\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}} + \frac{\partial^2 \ln p(\mathbf{y},\tilde{\boldsymbol{\theta}}_{m2})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML}),$$

where $\tilde{\boldsymbol{\theta}}_{m2}$ lies on the segment between $\hat{\boldsymbol{\theta}}_m$ and $\hat{\boldsymbol{\theta}}_{ML}$. Thus,

$$\begin{aligned}
\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML} &= \left[\frac{\partial^2 \ln p(\mathbf{y},\tilde{\boldsymbol{\theta}}_{m2})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right]^{-1}\frac{\partial \ln p(\mathbf{y},\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}} \\
&= L_n^{-(2)}\left(\tilde{\boldsymbol{\theta}}_{m2}\right)\left[\frac{\partial \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}} + \frac{\partial \ln p(\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}}\right] \\
&= L_n^{-(2)}\left(\tilde{\boldsymbol{\theta}}_{m2}\right)\left[0 + \frac{\partial \ln p(\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}}\right] \\
&= O_p(n^{-1})O_p(1) = O_p(n^{-1}).
\end{aligned}$$

Again, using the Taylor expansion, we can get

$$\begin{aligned}
&\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_m) - \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML}) \\
=\ & \frac{\partial \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{ML})}{\partial\theta}(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML}) + \frac{1}{2}(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML})'\frac{\partial^2 \ln p(\mathbf{y}|\tilde{\boldsymbol{\theta}}_{m3})}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}(\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML}) \\
=\ & O_p(n^{\frac{1}{2}})O_p(n^{-1}) + O_p(n^{-1})O_p(n)O_p(n^{-1}) = O_p(n^{-\frac{1}{2}}),
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{m3}$ lies on the segment between $\hat{\boldsymbol{\theta}}_m$ and $\hat{\boldsymbol{\theta}}_{ML}$. Therefore, we can have

$$\begin{aligned}
&-2\ln p(\mathbf{y}_{rep}|\mathbf{y}) = -2\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - 2\ln(1+c_1+c_2) \\
&= -2\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) - 2\ln\left[1+o_p(1)\right] = -2\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{m}}) + o_p(1) \\
&= -2\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + O_p(n^{-\frac{1}{2}}) + o_p(1) \\
&= -2\ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + o_p(1)
\end{aligned}$$

According to the derivation of AIC in Burnham and Anderson (2002), we have

$$\begin{aligned} & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[\mathcal{L}(\mathbf{y}_{rep}, \mathbf{y})\right] = E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p(\mathbf{y}_{rep}|\mathbf{y})\right] \\ = & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + o_p(1)\right] \\ = & E_{\mathbf{y}} \left[-2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{\boldsymbol{ML}}) + 2P + o_p(1)\right] \end{aligned}$$

In light of Lemma 3.1, using the Taylor expansion, we get

$$\begin{aligned} & \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) = \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) + \frac{\partial \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)}{\partial \theta}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m) + \frac{1}{2}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m)' \frac{\partial^2 \ln p(\mathbf{y}|\tilde{\boldsymbol{\theta}}_{m4})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m) \\ = & \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) + O_p(n^{1/2}) o_p(n^{-1/2}) + o_p(n^{-1/2}) O_p(n) o_p(n^{-1/2}) \\ = & \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) + o_p(1) = \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML}) + o_p(1), \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{m4}$ lies on the segment between $\bar{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{\theta}}_m$.

Hence, according to the proof of Lemma 3.1, we can get

$$\begin{aligned} & P_D = \int -2 \left[\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})\right] p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} \\ = & \int -2 \left[\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)\right] p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} + 2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m) - 2 \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) \\ = & -2 \frac{\partial \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)}{\partial \theta}(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m) - \int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' \frac{\partial^2 \ln p(\mathbf{y}|\tilde{\boldsymbol{\theta}}_{m5})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} + o_p(1) \\ = & o_p(1) - \int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\tilde{\boldsymbol{\theta}}_{m5})(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} + \int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' \left[\frac{\partial \ln p(\tilde{\boldsymbol{\theta}}_{m5})}{\partial \boldsymbol{\theta}\boldsymbol{\theta}'}\right] (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} \\ = & -\int (\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m)' L_n^{(2)}(\hat{\boldsymbol{\theta}}_m)(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}_m) p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta} + o_p(1) + O_p(n^{-1}) \\ = & -\mathbf{tr}\left\{L_n^2(\hat{\boldsymbol{\theta}}_m) V(\hat{\boldsymbol{\theta}}_m)\right\} + o_p(1) \\ = & \mathbf{tr}\left\{L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) \left[L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1})\right]\right\} \\ = & \mathbf{tr}\left[L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m)\right] + \mathbf{tr}\left[L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) o_p(n^{-1})\right] \\ = & P + o_p(1), \end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_{m5}$ lies on the segment between $\boldsymbol{\theta}$ and $\hat{\boldsymbol{\theta}}_m$.

Finally, we have

$$\begin{aligned} & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[\mathcal{L}(\mathbf{y}_{rep}, \mathbf{y})\right] \\ = & E_{\mathbf{y}} \left[-2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML}) + 2P + o_p(1)\right] \\ = & E_{\mathbf{y}} \left[-2 \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) + o_p(1) + 2P + o_p(1)\right] \\ = & E_{\mathbf{y}} \left[D(\bar{\boldsymbol{\theta}}) + 2P_D + o_p(1)\right] \\ = & E_{\mathbf{y}} \left[\mathrm{DIC}_1 + o_p(1)\right]. \end{aligned}$$

**Proof of Theorem 4.1**

According to Lemma 3.1, it can be shown that

$$\begin{aligned}
& V(\bar{\boldsymbol{\theta}}) = E\left[(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}})'|\mathbf{y}\right] = E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m+\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m+\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})'|\mathbf{y}\right] \\
= & E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)'|\mathbf{y}\right] + 2E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)|\mathbf{y}\right](\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})' + (\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})' \\
= & E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)'|\mathbf{y}\right] + 2(\bar{\boldsymbol{\theta}}-\hat{\boldsymbol{\theta}}_m)(\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})' + (\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})' \\
= & E\left[(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}_m)'|\mathbf{y}\right] - (\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_m-\bar{\boldsymbol{\theta}})' \\
= & V(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1/2})o_p(n^{-1/2}) \\
= & V(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1}) \\
= & L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1})
\end{aligned}$$

According to the proof of Theorem 3.1, it can be shown that

$$\hat{\boldsymbol{\theta}}_m - \hat{\boldsymbol{\theta}}_{ML} = O_p(n^{-1})$$

Then, based on the standard ML large sample theory and Lemma 3.1, we have $\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}}_m + o_p(n^{-1/2})$ and $\hat{\boldsymbol{\theta}}_{ML} = \boldsymbol{\theta}_0 + o_p(1)$ so that $\hat{\boldsymbol{\theta}}_m = \boldsymbol{\theta}_0 + o_p(1)$ and $\bar{\boldsymbol{\theta}} = \boldsymbol{\theta}_0 + o_p(1)$. Hence, $\hat{\boldsymbol{\theta}}_{ML}$, and $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$ both consistent estimators of $\boldsymbol{\theta}_0$. The standard ML theory and Assumption 6 suggest that

$$\begin{aligned}
\frac{1}{n}\mathbf{I}(\bar{\boldsymbol{\theta}}) &= -\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{H}(\boldsymbol{\theta}_0) + o_p(1), \\
\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\theta}}_m) &= -\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_m)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{H}(\boldsymbol{\theta}_0) + o_p(1) \\
\frac{1}{n}\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) &= -\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{H}(\boldsymbol{\theta}_0) + o_p(1)
\end{aligned}$$

where $\mathbf{H}(\boldsymbol{\theta})$ is the Hessian information matrix given by

$$\mathbf{H}(\boldsymbol{\theta}) = \int -\frac{1}{n}\left[\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right] g(\mathbf{y})\mathrm{d}\mathbf{y} = 0.$$

Then, we can get

$$\mathbf{I}(\bar{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}}_m) + o_p(n) = \mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)$$

Furthermore, since $p(\boldsymbol{\theta}) = O_p(1)$, it is noted that

$$L_n^{(2)}(\hat{\boldsymbol{\theta}}_m) = \mathbf{I}(\hat{\boldsymbol{\theta}}_m) + \frac{\partial^2 \ln p(\hat{\boldsymbol{\theta}}_m)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \mathbf{I}(\hat{\boldsymbol{\theta}}_m) + O_p(1) = \mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)$$

Hence, according to the proof of Theorem 3.1, we get

$$\begin{aligned}
P_D^* &= \mathbf{tr}\left[\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right] = \mathbf{tr}\left\{\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)\right]\left[V(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1})\right]\right\} \\
&= \mathbf{tr}\left\{\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)\right]\left[L_n^{-(2)}(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1})\right]\right\} \\
&= \mathbf{tr}\left\{\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)\right]\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_m) + o_p(n^{-1})\right]\right\} \\
&= \mathbf{tr}\left\{\left[\mathbf{I}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n)\right]\left[\mathbf{I}^{-1}(\hat{\boldsymbol{\theta}}_{ML}) + o_p(n^{-1})\right]\right\} \\
&= P + o_p(1) = P_D + o_p(1)
\end{aligned}$$

Hence,

$$\mathrm{RDIC} = \mathrm{DIC}_1 + o_p(1).$$

## Proof of Corollary 4.3

By the Talyor expansion, we get

$$\begin{aligned}
-\frac{1}{n}I\left(\overline{\boldsymbol{\theta}}\right) &= \nabla\psi_n\left(\bar{\boldsymbol{\theta}}\right) = \nabla\psi_n\left(\hat{\boldsymbol{\theta}}_m\right) + \nabla^2\psi_n\left(\hat{\boldsymbol{\theta}}_m\right)\left[I_P \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right)\right] \\
&\quad + \frac{1}{2}\nabla^3\psi_n\left(\widetilde{\boldsymbol{\theta}}\right)\left[I_P \otimes \left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}_m\right) \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right)\right],
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}$ lies between $\hat{\boldsymbol{\theta}}_m$ and $\bar{\boldsymbol{\theta}}$. Hence,

$$\begin{aligned}
I\left(\bar{\boldsymbol{\theta}}\right) &= I\left(\hat{\boldsymbol{\theta}}_m\right) - n\nabla^2\psi_n\left(\hat{\boldsymbol{\theta}}_m\right)\left[I_P \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right)\right] - \frac{1}{2}n\nabla^3\psi_n\left(\widetilde{\boldsymbol{\theta}}\right)\left[I_P \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right) \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right)\right] \\
&= I\left(\hat{\boldsymbol{\theta}}_m\right) - n\nabla^2\psi_n\left[I_P \otimes \left(\bar{\boldsymbol{\theta}} - \hat{\boldsymbol{\theta}}_m\right)\right] + o_p\left(1\right) \\
&= I\left(\hat{\boldsymbol{\theta}}_m\right) - n\nabla^2\psi_n\left[I_P \otimes \varrho_{-1}\right] + o_p\left(1\right) \\
&= I\left(\hat{\boldsymbol{\theta}}_m\right) - n\overline{H}_2\left[I_P \otimes \varrho_{-1}\right] + o_p\left(1\right),
\end{aligned}$$

$$\begin{aligned}
nL_n^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right) &= \left(\nabla\psi_n\left(\hat{\boldsymbol{\theta}}_m\right) + \frac{1}{n}\nabla\gamma^{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_m\right)\right)^{-1} = \left[\nabla\psi_n\left(\hat{\boldsymbol{\theta}}_m\right)\right]^{-1} + O_p\left(n^{-1}\right) \\
&= \left[\nabla^2\psi_n\right]^{-1} + O_p\left(n^{-1/2}\right) = \overline{H}_1^{-1} + O_p\left(n^{-1/2}\right) = Q + O_p\left(n^{-1/2}\right).
\end{aligned}$$

From Li et al. (2015), we have

$$V\left(\bar{\boldsymbol{\theta}}\right) = -L^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right) + \frac{1}{n^2}C(\hat{\boldsymbol{\theta}}_m) + o_p\left(n^{-2}\right),$$

where $C(\boldsymbol{\theta})$ is a bounded and continuous function of $\boldsymbol{\theta}$. Hence we get

$$\begin{aligned}
P_D^* &= tr\left[I\left(\bar{\boldsymbol{\theta}}\right)V\left(\bar{\boldsymbol{\theta}}\right)\right] = tr\left[I\left(\hat{\boldsymbol{\theta}}_m\right)V\left(\bar{\boldsymbol{\theta}}\right)\right] - tr\left(n\bar{H}_2\left[I_P\otimes\varrho_{-1}\right]V\left(\bar{\boldsymbol{\theta}}\right)\right) + o_p\left(n^{-1}\right) \\
&= tr\left[I\left(\hat{\boldsymbol{\theta}}_m\right)V\left(\bar{\boldsymbol{\theta}}\right)\right] - tr\left[\bar{H}_2\left[I_P\otimes\varrho_{-1}\right]\left(nV\left(\bar{\boldsymbol{\theta}}\right)\right)\right] + o_p\left(n^{-1}\right) \\
&= tr\left[\left(-L_n^{(2)}\left(\hat{\boldsymbol{\theta}}_m\right) - \left(-\nabla\gamma^{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\right)V\left(\bar{\boldsymbol{\theta}}\right)\right] - tr\left[\bar{H}_2\left[I_P\otimes\varrho_{-1}\right]\left(nV\left(\bar{\boldsymbol{\theta}}\right)\right)\right] + o_p\left(n^{-1}\right) \\
&= tr\left[\left(-L_n^{(2)}\left(\hat{\boldsymbol{\theta}}_m\right) - \left(-\nabla\gamma^{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\right)\left(-L_n^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right) + \frac{1}{n^2}C\left(\hat{\boldsymbol{\theta}}_m\right) + o_p\left(n^{-2}\right)\right)\right] \\
&\quad + tr\left[\bar{H}_2\left[I_P\otimes\varrho_{-1}\right]\left(-nL_n^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\right] + o_p(n^{-1}) \\
&= p - \frac{1}{n}tr\left[\left(-\nabla\gamma^{\boldsymbol{\theta}}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\left(-nL_n^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\right] + \frac{1}{n^2}tr\left[-L_n^{(2)}\left(\hat{\boldsymbol{\theta}}_m\right)C\left(\hat{\boldsymbol{\theta}}_m\right)\right] \\
&\quad + tr\left[\bar{H}_2\left[I_P\otimes\varrho_{-1}\right]\left(-nL_n^{(-2)}\left(\hat{\boldsymbol{\theta}}_m\right)\right)\right] + o_p(n^{-1}) \\
&= p - \frac{1}{n}tr\left[\left(-\nabla\gamma^{\boldsymbol{\theta}}\right)\left(-\overline{H}_1^{-1}\right)\right] + \frac{1}{n}tr\left[-\overline{H}_1\overline{C}\right] + tr\left[\overline{H}_2\left[I_P\otimes\varrho_{-1}\right]\left(-\overline{H}_1^{-1}\right)\right] + o_p\left(n^{-1}\right) \\
&= p - \frac{1}{n}tr\left[\nabla\gamma^{\boldsymbol{\theta}}\overline{H}_1^{-1}\right] - \frac{1}{n}tr\left[\overline{H}_1\bar{C}\right] - tr\left[\overline{H}_2\left(I_P\otimes\varrho_{-1}\right)\overline{H}_1^{-1}\right] + o_p\left(n^{-1}\right).
\end{aligned}$$

## The derivation of RDIC for the dynamic factor models

The complete-data log-likelihood function is:

$$\begin{aligned}
\ln f\left(Y,F|L,\Sigma,\Phi,Q\right) &= -\frac{\left(K+N\right)T-K}{2}\ln 2\pi - \frac{T}{2}\ln|\Sigma| - \frac{1}{2}\mathbf{tr}\left[\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\right] \\
&\quad -\frac{T-1}{2}\ln|Q| - \frac{1}{2}\mathbf{tr}\left[Q^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)\right],
\end{aligned}$$

where $Y=[Y_1',Y_2',...,Y_T']'$, $F=[F_1',F_2',...,F_T']'$, $F_{+1}=[F_2',F_3',...,F_T']'$, $F_{-1}=\left[F_1',F_2',...,F_{T-1}'\right]'$. Denote this function by $\varphi(L,\Sigma,\Phi,Q)$, In this appendix, we derive the first and second derivative of the complete-data log-likelihood function. The matrix differentiation used here follows the rules discussed in Magnus and Neudecker (1999).

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

Whenever there is no confusion, we denote $\varphi(L,\Sigma,\Phi,Q)$ simply by $\varphi$. The differential of

$\varphi(L,\Sigma,\Phi,Q)$ with respect to $L$ is

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

where

$$\widetilde{c}=\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right).$$

Taking $vec$ both sides, we get

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

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

$$D_L\left(\varphi\right)=\left(vec\left(\left[\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)\right]'\right)\right)'.$$

Similarly, we have

$$D_\Sigma\left(\varphi\right)=\left(vec\left(-\frac{T}{2}\Sigma^{-1}+\frac{1}{2}\Sigma^{-1}\left(Y-FL'\right)'\left(Y-FL'\right)\Sigma^{-1}\right)'\right)',$$

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

$$D_Q\left(\varphi\right)=\left(vec\left(-\frac{T-1}{2}Q^{-1}+\frac{1}{2}Q^{-1}\left(F_{+1}-F_{-1}\Phi'\right)'\left(F_{+1}-F_{-1}\Phi'\right)Q^{-1}\right)'\right)'.$$

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

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

$$d\widetilde{c}=d\left(\frac{1}{2}F'\left(Y-FL'\right)\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right)\right)=-\frac{1}{2}F'F\left(dL\right)'\left(\left(\Sigma^{-1}\right)'+\Sigma^{-1}\right).$$

And the second order derivative is

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

Then, we have,

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

$$\begin{aligned}
H &= G(T) = T', \quad T = S(\Sigma) = \frac{1}{2}F'\left(Y - FL'\right)\left(\left(\Sigma^{-1}\right)' + \Sigma^{-1}\right), \\
D(G(T)) &= K_K, \\
D(S(\Sigma)) &= I_N \otimes \left(F'\left(Y - FL'\right)\right) \cdot \left(-\frac{1}{2}\left(K_{NN} + I_{NN}\right)\right) \cdot \left(\left(\Sigma^{-1}\right)' \otimes \Sigma^{-1}\right), \\
DH(\Sigma) &= (DG(T))(DS(\Sigma)),
\end{aligned}$$

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

$$\begin{aligned}
D_{L,\Sigma}(\varphi) &= \frac{\partial D_L(\varphi)}{(\partial vec\Sigma)'} = (DG(T))(DS(\Sigma)) \\
&= K_{KN} \cdot I_N \otimes \left(F'\left(Y - FL'\right)\right) \cdot \left(-\frac{1}{2}\left(K_{NN} + I_{NN}\right)\right) \cdot \left(\left(\Sigma^{-1}\right)' \otimes \Sigma^{-1}\right),
\end{aligned}$$

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

$$D_{\Sigma,\Sigma}(\varphi) = K_{NN} \cdot \begin{pmatrix} \frac{T}{2} \cdot \frac{1}{2}\left(\left(\Sigma^{-1}\right)' \otimes \Sigma^{-1} + \left(\Sigma^{-1}\right)' \otimes \Sigma^{-1}\right) \\ -\frac{1}{2}\begin{pmatrix} \left(\Sigma^{-1}(Y - FL')'(Y - FL')\Sigma^{-1}\right)' \otimes \Sigma^{-1} \\ +\left(\Sigma^{-1}\right)' \otimes \left(\Sigma^{-1}(Y - FL')'(Y - FL')\Sigma^{-1}\right) \end{pmatrix} \end{pmatrix},$$

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

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

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

$$D_{Q,Q}(\varphi) = K_{KK} \cdot \begin{pmatrix} \frac{T-1}{2} \cdot \frac{1}{2}\left(\left(Q^{-1}\right)' \otimes Q^{-1} + \left(Q^{-1}\right)' \otimes Q^{-1}\right) \\ -\frac{1}{2}\begin{pmatrix} \left(Q^{-1}\left(F_{+1} - F_{-1}\Phi'\right)'\left(F_{+1} - F_{-1}\Phi'\right)Q^{-1}\right)' \otimes Q^{-1} \\ +\left(Q^{-1}\right)' \otimes \left(\Sigma^{-1}\left(F_{+1} - F_{-1}\Phi'\right)'\left(F_{+1} - F_{-1}\Phi'\right)Q^{-1}\right) \end{pmatrix} \end{pmatrix}.$$

**The special structure of parameter matrix:**

Let $L, \Sigma, \Phi, Q$ have some special structures. In particular, let

$$L^* = vec\left(\bar{L}\right),$$

where $\bar{L}$ is the last $(N-K)\times K$ block of $L$, and

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

**The first order derivatives are as follows:**

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

**The second order derivatives are as follows:**

$$\begin{aligned}
D_{L^{*},L^{*}}\left(\varphi\right) &= D_{L^{*}}\left(D_{L^{*}}\left(\varphi\right)\right)=D_{L^{*}}\left(D_{L}\left(\varphi\right)\cdot\dot{I}_{L^{*}}\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{L^{*}}\left(D_{L}\left(\varphi\right)\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{L,L}\left(\varphi\right)\cdot\dot{I}_{L^{*}},\\
D_{L^{*},\Sigma^{*}}\left(\varphi\right) &= D_{\Sigma^{*}}\left(D_{L^{*}}\left(\varphi\right)\right)=D_{\Sigma^{*}}\left(D_{L}\left(\varphi\right)\cdot\dot{I}_{L^{*}}\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{\Sigma^{*}}\left(D_{L}\left(\varphi\right)\right)\\
&= \left(\dot{I}_{L^{*}}^{'}\otimes I_{1}\right)\cdot D_{\Sigma}\left(D_{L}\left(\varphi\right)\right)\cdot D_{\Sigma^{*}}\left(\Sigma\left(\Sigma^{*}\right)\right)\\
&= \dot{I}_{L^{*}}^{'}\cdot D_{L,\Sigma}\left(\varphi\right)\cdot\dot{I}_{\Sigma^{*}},\\
D_{L^{*},\Phi^{*}}\left(\varphi\right) &= 0,\\
D_{L^{*},Q^{*}}\left(\varphi\right) &= 0,
\end{aligned}$$

$$\begin{aligned}
D_{\Sigma^{*},\Sigma^{*}}\left(\varphi\right) &= D_{\Sigma^{*}}\left(D_{\Sigma^{*}}\left(\varphi\right)\right)=D_{\Sigma^{*}}\left(D_{\Sigma}\left(\varphi\right)\cdot\dot{I}_{\Sigma^{*}}\right)\\
&= \dot{I}_{\Sigma^{*}}^{'}\otimes I_{1}\cdot D_{\Sigma^{*}}\left(D_{\Sigma}\left(\varphi\right)\right)\\
&= \dot{I}_{\Sigma^{*}}^{'}\cdot D_{\Sigma}\left(D_{\Sigma}\left(\varphi\right)\right)\cdot\dot{I}_{\Sigma^{*}},\\
D_{\Sigma^{*},\Phi^{*}}\left(\varphi\right) &= 0,\\
D_{\Sigma^{*},Q^{*}}\left(\varphi\right) &= 0.
\end{aligned}$$

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

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

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

$$\dot{I}_{L^{*}}=diag\left(P_{1},P_{2},...,P_{K}\right),$$

where

$$P_i = \left[ \begin{array}{c} 0_{K\times(N-K)} \\ I_{N-K} \end{array} \right].$$

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

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

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

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

where

$$R_k = \left[ \begin{array}{c} 0_{(k-1)\times(K-k+1)} \\ I_{K-k+1} \end{array} \right]_{K\times(K-k+1)},$$

since $Q$ is a symmetric matrix.

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

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

The second order derivatives of the complete-data likelihood with respect to $L^*, \Sigma^*, \Phi^*, Q^*$ are:

$$\left[ \begin{array}{cccc} D_{L^*,L^*}(\varphi) & D_{L^*,\Sigma^*}(\varphi) & 0 & 0 \\ D_{\Sigma^*,L^*}(\varphi) & D_{\Sigma^*,\Sigma^*}(\varphi) & 0 & 0 \\ 0 & 0 & D_{\Phi^*,\Phi^*}(\varphi) & D_{\Phi^*,Q^*}(\varphi) \\ 0 & 0 & D_{Q^*,\Phi^*}(\varphi) & D_{Q^*,Q^*}(\varphi) \end{array} \right].$$

**The derivation of RDIC for the stochastic volatility models**

**The derivatives of the complete-data log-likelihood for $M_1$**

**The complete-data log-likelihood function**

$$\begin{aligned} \ln p(\mathbf{y}, \mathbf{h}|\boldsymbol{\theta}) &= -n\ln 2\pi + \frac{n}{2}\ln\nu - \frac{1}{2}\sum_{t=1}^{n} h_t - \frac{1}{2}\sum_{t=1}^{n}\frac{(y_t-\alpha)^2}{\exp(h_t)} \\ &\quad -\frac{1}{2}\nu\left[\sum_{t=1}^{n}\left(h_t - \mu - \phi\left(h_{t-1}-\mu\right)\right)^2\right], \end{aligned}$$

where $\mathbf{h} = (h_1, h_2, ..., h_n)'$, $\nu = 1/\tau^2$.

**The first order derivatives**

$$\frac{\partial \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \alpha}=\sum_{t=1}^{n}\frac{\left(y_t-\alpha\right)}{\exp\left(h_t\right)},$$

$$\begin{aligned}\frac{\partial \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \mu} &= -\frac{1}{2}\nu\left[-2\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)\left(1-\phi\right)\right]\\ &= \nu\left[\left(1-\phi\right)\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \phi} &= -\frac{1}{2}\nu\left[-2\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)\left(h_{t-1}-\mu\right)\right]\\ &= \nu\left[\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)\left(h_{t-1}-\mu\right)\right],\end{aligned}$$

$$\frac{\partial \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \nu}=\frac{n}{2}\frac{1}{\nu}-\frac{1}{2}\left[\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)^2\right].$$

**The second order derivatives**

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \alpha\partial \alpha}=-\sum_{t=1}^{n}\frac{1}{\exp\left(h_t\right)}=-\sum_{t=1}^{n}\exp\left(-h_t\right),$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \alpha\partial \mu}=\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \alpha\partial \phi}=\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \alpha\partial \nu}=0,$$

$$\begin{aligned}\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \mu\partial \mu} &= \nu\left[-\left(1-\phi\right)\sum_{t=1}^{n}\left(1-\phi\right)\right]\\ &= -\nu\left[n\left(1-\phi\right)^2\right],\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \mu\partial \phi} &= \nu\left[-\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)-\left(1-\phi\right)\sum_{t=1}^{n}\left(h_{t-1}-\mu\right)\right]\\ &= -\nu\left[\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)+\left(1-\phi\right)\sum_{t=1}^{n}\left(h_{t-1}-\mu\right)\right],\end{aligned}$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \mu\partial \nu}=\left(1-\phi\right)\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right),$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \phi\partial \phi}=\nu\left[-\sum_{t=1}^{n}\left(h_{t-1}-\mu\right)^2\right],$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \phi\partial \nu}=\sum_{t=1}^{n}\left(h_t-\mu-\phi\left(h_{t-1}-\mu\right)\right)\left(h_{t-1}-\mu\right),$$

$$\frac{\partial^2 \ln p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)}{\partial \nu\partial \nu}=-\frac{n}{2\nu^2}.$$

**The derivatives of the complete-data log-likelihood for $M_2$**

**The complete-data log-likelihood function**

$$\begin{aligned}\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right) &= -\frac{n}{2}\ln 2\pi+\frac{n}{2}\ln\nu-\frac{1}{2}\nu\left[\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)^2\right]\\ &\quad -\frac{1}{2}\sum_{t=1}^{n}\frac{\left(y_t-\alpha\right)^2}{\sigma_t^2}-\frac{n}{2}\ln 2\pi-\frac{1}{2}\sum_{t=1}^{n}\sigma_t^2,\end{aligned}$$

where $\boldsymbol{\sigma^2}=\left(\sigma_1^2,\sigma_2^2,...\sigma_n^2\right)'$.

**The first order derivatives**

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\alpha}=\sum_{t=1}^{n}\frac{y_t-\alpha}{\sigma_t^2},$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\mu}=\nu\left[\left(1-\phi\right)\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\phi}=\nu\left[\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\left(\ln\sigma_{t-1}^2-\mu\right)\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\nu}=\frac{n}{2\nu}-\frac{1}{2}\left[\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)^2\right].$$

**The second order derivatives**

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\alpha\partial\alpha}=-\sum_{t=1}^{n}\frac{1}{\sigma_t^2},$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\alpha\partial\mu}=\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\alpha\partial\phi}=\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\alpha\partial\nu}=0,$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\mu\partial\mu}=-\nu\left[n\left(1-\phi\right)^2\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\mu\partial\phi}=-\nu\left[\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)+\left(1-\phi\right)\sum_{t=1}^{n}\left(\ln\sigma_{t-1}^2-\mu\right)\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\mu\partial\nu}=\left(1-\phi\right)\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right),$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\phi\partial\phi}=\nu\left[-\sum_{t=1}^{n}\left(\ln\sigma_{t-1}^2-\mu\right)^2\right],$$

$$\frac{\partial\ln p\left(\mathbf{y},\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\phi\partial\nu}=\sum_{t=1}^{n}\left(\ln\sigma_t^2-\mu-\phi\left(\ln\sigma_{t-1}^2-\mu\right)\right)\left(\ln\sigma_{t-1}^2-\mu\right),$$

$$\frac{\partial^2\ln p\left(\mathbf{y},\mathbf{ln}\,\boldsymbol{\sigma^2}|\boldsymbol{\theta}\right)}{\partial\nu\partial\nu}=-\frac{n}{2\nu^2}.$$

**Gaussian Approximation**

The complete-data log-likelihood function of $M_1$ can be also expressed as:

$$\begin{aligned}\ln\left(p\left(\mathbf{y},\mathbf{h}|\boldsymbol{\theta}\right)\right) &= -\frac{n}{2}\ln\left(2\pi\right)-\frac{n}{2}\ln\left(\tau^2\right)-\frac{1}{2}\left(\mathbf{h}-\boldsymbol{\mu}\right)'\boldsymbol{Q}\left(\mathbf{h}-\boldsymbol{\mu}\right)\\ &\quad -\frac{n}{2}\ln\left(2\pi\right)-\frac{1}{2}\sum_{t=1}^{n}h_t-\sum_{t=1}^{n}\frac{\left(y_t-\alpha\right)^2}{2}\exp\left(-h_t\right),\end{aligned}$$

where $\boldsymbol{\mu}=\mu\mathbf{e}$, $\mathbf{e}'=(1,\ldots,1)_n$, $\boldsymbol{Q}$ is a tri-diagonal precision matrix, $\boldsymbol{Q}=\boldsymbol{Q}^*/\tau^2$, $\boldsymbol{Q}^*$ is defined as follows:

$$\boldsymbol{Q}^*=\begin{pmatrix}\phi^2 & -\phi & & & & \\ -\phi & 1+\phi^2 & -\phi & & & \\ & & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot & \\ & & & -\phi & 1+\phi^2 & -\phi \\ & & & & -\phi & 1\end{pmatrix}.$$

The posterior density of $\mathbf{h}$ is

$$\begin{aligned}p\left(\mathbf{h}|\mathbf{y},\boldsymbol{\theta}\right) &\propto \exp\left[-\frac{1}{2}\left(\mathbf{h}-\boldsymbol{\mu}\right)'\boldsymbol{Q}\left(\mathbf{h}-\boldsymbol{\mu}\right)-\sum_{t=1}^{n}\left(\frac{1}{2}h_t+\frac{\left(y_t-\alpha\right)^2}{2}\exp\left(-h_t\right)\right)\right]\\ &= \exp\left(f\left(\mathbf{h}\right)\right)\approx\exp\left(-\frac{1}{2}\mathbf{h}'\boldsymbol{c}\mathbf{h}+\boldsymbol{b}\mathbf{h}+\text{constant}\right).\end{aligned}$$

To obtain the parameters $\mathbf{c}$ and $\mathbf{b}$ of the canonical form, we use the first and second order derivatives:

$$\begin{aligned}\dot{f}\left(\mathbf{h}\right) &= -\mathbf{h}'\boldsymbol{Q}+\boldsymbol{\mu}'\boldsymbol{Q}-\frac{1}{2}\mathbf{e}'+\frac{1}{2}\left(\mathbf{y}^{*2}\right)'\odot\exp\left(-\mathbf{h}\right)'\\ \ddot{f}\left(\mathbf{h}\right) &= -\boldsymbol{Q}-diag\left(\frac{1}{2}\left(\mathbf{y}^*\right)^2\odot\exp\left(-\mathbf{h}\right)\right),\end{aligned}$$

where $\mathbf{y}^*=\mathbf{y}-\boldsymbol{\alpha}$ and $\boldsymbol{\alpha}=\alpha\mathbf{e}$, $\mathbf{e}'=(1,\ldots,1)_n$, $\mathbf{y}^{*2}=(y_1^{*2},\ldots,y_n^{*2})'$ and $\exp(-\mathbf{h})=(\exp(-h_1),\ldots,\exp(-h_n))'$.

Denoting the mode of $f$ by $\mathbf{m}$, we apply the Taylor expansion to $f\left(x\right)$:

$$\begin{aligned}f\left(\mathbf{h}\right) &\approx \left(\mathbf{h}-\mathbf{m}\right)'\frac{\ddot{f}\left(\mathbf{m}\right)}{2}\left(\mathbf{h}-\mathbf{m}\right)+\dot{f}\left(\mathbf{m}\right)\left(\mathbf{h}-\mathbf{m}\right)+\text{constant}\\ &= -\frac{1}{2}\mathbf{h}'\left(-\ddot{f}\left(\mathbf{m}\right)\right)\mathbf{h}-\mathbf{m}'\ddot{f}\left(\mathbf{m}\right)\mathbf{h}+\dot{f}\left(\mathbf{m}\right)\mathbf{h}+\text{constant}\\ &= -\frac{1}{2}\mathbf{h}'\mathbf{c}\mathbf{h}+\mathbf{b}\mathbf{h}+\text{constant}.\end{aligned}$$

Now, we obtain $\mathbf{c}$ and $\mathbf{b}$ as

$$\mathbf{c}=-\ddot{f}\left(\mathbf{m}\right)=\mathbf{Q}+diag\left(\frac{1}{2}\mathbf{y}^{*2}\odot\exp\left(-\mathbf{m}\right)\right),$$

$$
\begin{aligned}
\mathbf{b} &= -\mathbf{m}'\ddot{f}(\mathbf{m}) + \dot{f}(\mathbf{m}) \\
&= \mathbf{m}'\mathbf{Q} + \mathbf{m}' diag\left(\frac{1}{2}\mathbf{y}^{*2} \odot \exp(-\mathbf{m})\right) \\
&\quad -\mathbf{m}'\mathbf{Q} + \boldsymbol{\mu}'\mathbf{Q} - \frac{1}{2}\mathbf{e}' + \frac{1}{2}\left(\mathbf{y}^{*2}\right)' \odot \exp(-\mathbf{m})' \\
&= \mathbf{m}' diag\left(\frac{1}{2}\mathbf{y}^{*2} \odot \exp(-\mathbf{m})\right) + \frac{1}{2}\left(\mathbf{y}^{*2}\right)' \odot \exp(-\mathbf{m})' + \boldsymbol{\mu}'\mathbf{Q} - \frac{1}{2}\mathbf{e}'.
\end{aligned}
$$

Using

$$
-\frac{1}{2}\mathbf{h}'\mathbf{c}\mathbf{h} + \mathbf{b}\mathbf{h} + \text{constant} = -\frac{1}{2}(\mathbf{h} - \mathbf{m}^*)'\mathbf{Q}^*(\mathbf{h} - \mathbf{m}^*),
$$

we obtain

$$
\begin{aligned}
\mathbf{Q}^* &= \mathbf{c} = \mathbf{Q} + diag\left(\frac{1}{2}\mathbf{y}^{*2} \odot \exp(-\mathbf{m})\right), \\
\mathbf{m}^* &= \mathbf{Q}^{*-1}\mathbf{b}'.
\end{aligned}
$$

To obtain the optimal mode of $\mathbf{Q}^*$ and $\mathbf{m}^*$, we run the above procedure recursively until convergence.

# References


**1** Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory*, Springer Verlag, **1**, 267-281.

**2** An, S. and Schorfheide, F. (2007). Bayesian analysis of DSGE models. *Econometric Reviews*, **26**, 113–172.

**3** Ando, T. (2007). Bayesian predictive information criterion for the evaluation of hierarchical Bayesian and empirical Bayes models. *Biometrika*, **94**, 443–458.

**4** Ando, T. and Tsay, R. (2010). Predictive likelihood for Bayesian model selection and averaging. *International Journal of Forecasting*, **26**, 744–763.

**5** Berg, A., Meyer, R. and Yu, J. (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics*, **22**, 107-120.

**6** Bester, C.A. and Hansen, C. (2006). Bias reduction for Bayesian and frequentist estimators. SSRN Working Paper Series

**7** Berger, J. O., Liseo, B., and Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science*, **14**, 1-28.

**8** Bernanke, B., Boivin, J. and Eliasz, P. (2005). Measuring the effects of monetary policy: a factor-augmented vector autoregressive (FAVAR) approach. *The Quarterly Journal of Economics*, **120**, 387-422.

**9** Black, F. (1976). Studies of stock market volatility changes. *Proceedings of the American Statistical Association, Business and Economic Statistics Section*, 177–181.

**10** Brooks, S. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002). *Journal of the Royal Statistical Society Series B*, **64**, 616–618.

**11** Burnham, K. and Anderson, D. (2002). *Model Selection and Multi-model Inference: A Practical Information-theoretic Approach*, Springer.

**12** Celeux, G., Forbes, F. Robert, C. and Titterington, D. (2006). Deviance Information Criteria for Missing Data Models. *Bayesian Analysis*, **1**, 651–674.

**13** Chan, J. C. and Jeliazkov, I. (2009). Efficient simulation and integrated likelihood estimation in state space models. *International Journal of Mathematical Modelling and Numerical Optimization*, **1**, 101–120.

**14** Chen, C. (1985). On asymptotic normality of limiting density functions with Bayesian implications. *Journal of the Royal Statistical Society Series B*, **47**, 540–546.

**15** Chib, S. (1995). Marginal Likelihood from the Gibbs Output. *Journal of the American Statistical Association*, **90**, 1313-1321.

**16** Chib, S. and Jeliazkov, I. (2001). Marginal likelihood from the Metropolis-Hastings output. *Journal of the American Statistical Association*, **96**, 270-281.

**17** Claeskens, G., and Hjort, N. L. (2003). The focused information criterion. *Journal of the American Statistical Association*, **98**, 900-916.

**18** Clark, P. (1973). A subordinated stochastic process model with finite variance for speculative prices. *Econometrica*, **41**, 135-155.

**19** Dejong, D. and Dave, C. (2007). *Structural Macroeconomics,* Princeton University Press, Princeton.

**20** Dempster, A., Laird, N. and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B*, **39**, 1-38.

**21** Gelfand, A. and Trevisani, M. (2002). Discussion on the paper by Spiegelhalter, Best, Carlin, and van de Linde (2002). *Journal of the Royal Statistical Society Series B*, **64**, 629-630.

**22** Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis.* CRC press.

**23** Geweke, J. (1977). The dynamic factor analysis of economic time-series models. *Latent Variables in Socio-economic Models*, ed. by A. Aigner and A. Goldberger, North-Holland, 365-395.

**24** Geweke, J. (2005). *Contemporary Bayesian Econometrics and Statistics.* Wiley-Interscience.

**25** Geweke, J., Koop, G. and van Dijk, H. (2011). *Oxford Handbook of Bayesian Econometrics*, Oxford University Press.

**26** Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*, Springer Verlag.

**27** Giannone, D., Reichlin, L. and Sala, L. (2004). Monetary policy in real time. *NBER Macroeconomics Annual*, 161-200.

**28** Gourieroux C, Monfort A. and Trognon, A. (1984). Pseudo maximum likelihood methods: Theory. *Econometrica*, **52(3)**, 681-700.

**29** Hansen, B. E. (2005). Challenges for econometric model selection. *Econometric Theory*, **21(1)**, 60-68.

**30** Huang, S. and Yu, J. (2010). Bayesian analysis of structural credit risk models with microstructure noises. *Journal of Economic Dynamics and Control*, **34**, 2259-2272.

**31** Ibrahim, J., Zhu, H. and Tang, N.S. (2008). Model selection criteria for missing-data problems using the EM algorithm. *Journal of the American Statistical Association*, **103**, 1648–1658.

**32** Kass, R. and Raftery, A. (1995). Bayes factors. *Journal of the American Statistical Association*, **90**, 773-795.

**33** Kim, J. (1994). Bayesian asymptotic theory in a time series model with a possible nonstationary process. *Econometric Theory*, **10**, 764-773.

**34** Kim, J. (1998). Large sample properties of posterior densities, Bayesian information criterion and the likelihood principle in nonstationary time series models. *Econometrica*, **66**, 359-380.

**35** Kim, S., Shephard, N. and Chib, S., (1998). Stochastic volatility: Likelihood inference and comparison with ARCH models. *The Review of Economic Studies*, **65**, 361-393.

**36** Kose, A. M., Otrok, C. and Whiteman, C. (2003). International business cycles: World, region, and country-specific factors. *American Economic Review*, **93**, 1216-1239.

**37** Kose, A.M., Otrok, C. and Whiteman, C. (2008). Understanding the evolution of world business cycles. *Journal of International Economics*, **75**, 110–130.

**38** Lancaster, T. (2000). The incidental parameter problem since 1948. *Journal of Econometrics*, **95(2)**, 391-413.

**39** Li, Y., Liu, X.B., Zeng, T., Yu, J., (2015), A Bayesian Wald test for hypothesis testing, Working Paper, Singapore Management University.

**40** Louis, T. (1982). Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society Series B*, **44**, 226–233.

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

**42** McLachlan, G. and Krishnan, T. (2008). *The EM Algorithm and Extensions*, John Wiley and Sons.

**43** Meyer, R. and Yu, J. (2000). BUGS for a Bayesian analysis of stochastic volatility models. *The Econometrics Journal*, **3**, 198-215.

**44** Newton, M. and Raftery, A. (1994). Approximate Bayesian inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society. Series B*, **56**, 3-48.

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

**46** Oakes, D. (1999). Direct calculation of the information matrix via the EM. *Journal of the Royal Statistical Society Series B*, **61**, 479-482.

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

**48** Plummer, M. (2006). Comment on Article by Celeux, et al. *Bayesian analysis*, **4(1)**, 681-686.

**49** Phillips, P.B.C. (1996). Econometric model determination. *Econometrica*, **64**, 763-812.

**50** Phillips, P.B.C. and Ploberger, W. (1996). An asymptotic theory of Bayesian inference for time series. *Econometrica*, **64**, 381-412.

**51** Rilstone, P., Srivatsava, V.K., and Ullah, A. (1996). The second order bias and MSE of nolinear estimators . *Journal of Econometrics*, **75**, 369-395.

**53** Rue, H., Martino, S. and Chopin, N. (2009). Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the Royal Statistical Society Series B*, **71**, 319-392.

**53** Rue, H., Steinsland, I. and Erland, S. (2004). Approximating hidden Gaussian Markov random fields. *Journal of the Royal Statistical Society Series B*, **66**, 877-892.

**54** Sargent, T. and Sims, C. (1977). Business cycle modeling without pretending to have too much a priori economic theory. *New Methods in Business Research, Federal Reserve Bank of Minneapolis.*

**55** Sawa, T. (1978). Information criteria for discriminating among alternative regression models. *Econometrica*, **46,** 1273-1291.

**56** Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B*, **64**, 583–639.

**57** Spiegelhalter, D., Best, N., Carlin, B. and van der Linde, A. (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society Series B*, **76**, 485-493.

**58** Stern, S. (1997). Simulation-based estimation. *Journal of Economic Literature*, **35**, 2006-2039.

**59** Stock, J. and Watson, M. (1999). Forecasting inflation. *Journal of Monetary Economics*, **44**, 293-335.

**60** Stock, J. and Watson, M. (2002). Macroeconomic Forecasting Using Diffusion Indexes. *Journal of Business and Economic Statistics*, **20**, 147-162.

**61** Stock, J. and Watson, M. (2011). Dynamic factor models. *Oxford Handbook of Economic Forecasting*, edited by M. P. Clements and D. F. Hendry, Oxford University Press.

**62** Takeuchi, K. (1976). Distribution of Informational Statistics and a Criterion of Model Fitting. *Suri-Kagaku (Mathematic Sciences)*, **153**, 12-18. (in Japanese).

**63** Tanner, M. and Wong, W.(1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, **82**, 528-540.

**64** Wu, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, **11**, 95-103.

**65** Verdinelli, I. and Wasserman, L. (1995). Computing Bayes factors using a generalization of the Savage-Dickey density ratio. *Journal of the American Statistical Association*, **90**, 614-618.

**66** Yu, J. (2005). On leverage in a stochastic volatility model. *Journal of Econometrics*, **127**, 165-178.