Integrated Deviance Information Criterion for Latent Variable Models[*]

Yong Li
Renmin University of China

Jun Yu
Singapore Management University

Tao Zeng
Zhejiang University

February 25, 2018

Abstract

Deviance information criterion (DIC) has been widely used for Bayesian model comparison, especially after Markov chain Monte Carlo (MCMC) is used to estimate candidate models. This paper studies the problem of using DIC to compare latent variable models after the models are estimated by MCMC together with the data augmentation technique. Our contributions are twofold. First, we show that when MCMC is used with data augmentation, it undermines theoretical underpinnings of DIC. As a result, by treating latent variables as parameters, the widely used way of constructing DIC based on the conditional likelihood, although facilitating computation, should not be used. Second, we propose two versions of integrated DIC (IDIC) to compare latent variable models without treating latent variables as parameters. The large sample properties of IDIC are studied and an asymptotic justification of IDIC is provided. Some popular algorithms such as the EM, Kalman and particle filtering algorithms are introduced to compute IDIC for latent variable models. IDIC is illustrated using asset pricing models, dynamic factor models, and stochastic volatility models.

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

[*] We wish to thank Eric Renault, 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. 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/. Tao Zeng, School of Economics, Zhejiang University, Zhejiang, China 310058. Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No. 71271221,71773130), Program for New Century Excellent Talents in University. Yu thanks the Singapore Ministry of Education for Academic Research Fund under grant number MOE2013-T3-1-009.

1 Introduction

Deviance information criterion (DIC) of Spiegelhalter, et al (2002) is a popular method for model selection in the Bayesian community. It has been used in a wide range of fields such as biostatistics, ecology, and economics. According to Spiegelhalter et al. (2014), Spiegelhalter, et al (2002) is the third most cited paper in international mathematical sciences between 1998 and 2008. Up to February 2018, it has received more than 5100 citations on the Web of Knowledge and nearly 9000 citations on Google Scholar.

The growth in popularity in DIC among applied researchers is understandable from a few aspects. First, DIC is a Bayesian version of the well-known Akaike Information Criterion (AIC) of Akaike (1973) that is based on maximum likelihood (ML). As shown in Li et al (2017), like AIC, DIC selects a model to minimize a plug-in predictive loss. This objective may appeal to applied researchers. Second, unlike AIC which is based on the log-likelihood function (or deviance) with the ML estimate (MLE) of parameters being plugged in, DIC is based on the deviance with the posterior mean of parameters being plugged in. Li, et al (2017) gives the details about the loss functions associated with AIC and DIC. The detach of DIC with ML is important when candidate models are difficult to estimate by ML. In this case, applied researchers may prefer Bayesian estimation methods over ML. In Bayesian statistics, the recent development of Markov chain Monte Carlo (MCMC) methods has been a key step in making it possible to estimate large hierarchical models. Large hierarchical models are typically difficult to estimate by ML, making ML-based model comparison criteria hard to implement. Third, DIC has a penalty term which can take account of prior information. This is different from AIC where the penalty term only depends on the number of parameters in a candidate model.

A typical hierarchical model used 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 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 factor model. On the basis of macroeconomic theory, the dynamic factor model attempts to explain aggregate economic phenomena by taking into account the fact that the economy is affected by some important factors. In microeconometrics, many discrete choice models and panel data models involve unobserved variables in order to capture observed heterogeneity across economic entities (Norets, 2009; Stern, 1997).

For latent variable models, the most popular approach to implementing MCMC is to employ the *data augmentation* strategy of Tanner and Wong (1987). This strategy expands the parameter space by treating latent variables as additional model parameters. Data augmentation greatly simplifies MCMC computation of posterior distributions and Bayesian estimation because it changes the likelihood function from observed-data likelihood to conditional likelihood (i.e. the likelihood conditional on the latent variables) which often has a closed-form expression. As DIC is based on the posterior mean of conditional likelihood, the closed-form expression of conditional likelihood greatly facilitates calculation of DIC for latent variable models. Not surprisingly, data augmentation has emerged as a standard method for implementing MCMC and for obtaining DIC for latent variable models. For example, it is the default choice if one uses WinBUGS (a popular Bayesian software). As acknowledged in Spiegelhalter, et al (2014), this default way of calculating DIC for latent variable models "is only to make the technique computationally feasible".

The first contribution of the present paper is that we show that the default way of calculating DIC for comparing latent variable models is asymptotically unjustifiable. The lack of justification arises because both the standard Bayesian large sample theory (such as the Bernstein–von Mises theorem) and the standard ML large sample theory (such as consistency and asymptotic normality) do not hold when latent variables are treated as parameters. As a result, the asymptotic theory developed in Li, et al (2017) is no longer applicable. In fact, the posterior distribution of latent variables may not be normally distributed as the sample size goes to infinity. The posterior means of latent variables may not be close to the MLE even asymptotically. Furthermore, as a practical problem, by expanding the parameter space, the data augmentation technique greatly increases the penalty term, making DIC very sensitive to apparently innocuous transformations and distributional representations of a candidate model.

Without using data augmentation, however, the (observed-data) likelihood function of many latent variable models does not have a closed-form expression. This is the exact reason why ML and hence AIC are difficult to use. Not surprisingly, DIC based on observed-data likelihood is also difficult to compute. As the second contribution of this paper, we introduce two new model selection criteria, which we call integrated DIC (IDIC), to make Bayesian comparison of latent variable models. Both of them are based on observed-data likelihood and the latent variables are not treated as parameters. One of them is constructed by using a plug-in predictive distribution while the other is by using the full Bayesian predictive distribution. Under some regularity conditions, the large sample properties of IDIC are studied. It is shown that the two versions of IDIC are asymptotically unbiased estimators of their respective risks which are the expected Kullback–Leibler divergence between the data generating process (DGP) and the respective predictive distributions. Hence, the two versions of IDIC select a model that asymptotically minimizes the respective risk.

The problem in DIC based on conditional likelihood has been pointed out in the literature. For example, Millar (2009) documented strong evidence of poor performance of DIC in negative binomial and Poisson-lognormal models using simulated data. He found that DIC almost always prefers the Poisson-gamma model instead of the Poisson-lognormal model, even when data are simulated from a Poisson-lognormal model. Millar and McKechnie (2014) documented strong evidence of poor performance of DIC in state-space models using simulated data. They further proposed a one-step-ahead DIC, where prediction is conditional on the state at the previous time point. Chan and Grant (2016a) showed that, in the context of stochastic volatility models, DIC tends to favor overfitted models using simulated data. They also showed that DIC based on observed-data likelihood performs well using simulated data. To compute DIC based on observed-data likelihood, they introduced an important-sampling-based algorithm. For three classes of latent variable models Chan and Grant (2016b) developed fast algorithms based on sparse matrix algorithms to compute observed-data-likelihood based DIC. However, none of these studies have provided any theoretical reason to show why conditional-likelihood based DIC is not justified and why the proposed solutions are asymptotically justified.

The paper is organized as follows. Section 2 reviews DIC for model comparison. In Section 3, we discuss several versions of DIC that exist in the literature for comparing latent variable models. We also explain why one of them is widely used and why this version of DIC is not theoretically justified. In Section 4, we introduce two versions of IDIC for comparing latent variable models. Large sample properties of IDIC are studied. Several general algorithms are introduced to compute IDIC in this section. Section 5 illustrates the method using three popular models in economics and finance, namely asset pricing models, dynamic factor models, stochastic volatility models. Section 6 concludes the paper. The Appendix collects the proof of the theoretical results in the paper.

2 DIC for Bayesian Model Comparison

Arguably the most important development in the Bayesian model comparison literature in recent years is DIC of Spiegelhalter, et al (2002). Compared with Bayes factors (BFs) which compare models through their "posterior probabilities" and try to search for the "true" model", DIC tries to find a better model for making prediction.

DIC enjoys several desirable features. First, DIC is easy to calculate when the likelihood function is available in closed-form and the posterior distributions of the models are obtained by MCMC. Second, it is applicable to a wide range of statistical models. Third, unlike BFs, it is not subject to the notorious Jeffreys-Lindley's paradox and can be used when noninformative or improper priors are used.

Consider a candidate parametric model, M, denoted by $p(\mathbf{y}|M, \boldsymbol{\theta})$ which is used to fit the data $\mathbf{y} := (y_1, y_2, \cdots, y_n)'$, where $\boldsymbol{\theta}$ is the parameter with P dimensions and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq R^P$. We will write $p(\mathbf{y}|M, \boldsymbol{\theta})$ as $p(\mathbf{y}|\boldsymbol{\theta})$ when there is no confusion. DIC of Spiegelhalter, et al (2002) is given by

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2P_D, \tag{1}$$

where $D(\boldsymbol{\theta}) = -2\ln p(\mathbf{y}|\boldsymbol{\theta})$, $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$, and P_D , which is known as "effective number of parameters", is given by:

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

DIC and AIC share a very important common feature, that is, they try to find a model that asymptotically minimizes the expected Kullback–Leibler divergence between the DGP and a predictive distribution; see Li, et al (2017). However, DIC and AIC have some important differences. First and foremost, AIC is based on the MLE while DIC is based on the posterior mean. The penalty term in DIC is determined by P_D whose value may depend on the prior. The penalty term in AIC depends on the number of parameters and hence it is invariant to the choice of priors.

As acknowledged in Spiegelhalter, et al (2002, 2014), the decision-theoretic justification of DIC is not rigorous in the literature. Very recently, under some mild regularity conditions, Li, et al (2017) provided a rigorous decision-theoretic justification to DIC when the standard Bayesian large sample theory and the standard ML large sample theory are valid. In this section, we first give a simple review of this justification for DIC.

Let $\mathbf{y}_{rep} = (y_{1,rep}, \cdots, y_{n,rep})$ be the independent replicate data of n observations generated by the same mechanism that gives rise to the observed data $\mathbf{y}$ and $g(\mathbf{y})$ is the DGP. 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 $p(\mathbf{y}_{rep}|\mathbf{y})$

$$\begin{aligned} KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] &= E_{\mathbf{y}_{rep}}\left[\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\mathbf{y}\right)}\right] \\ = & \int \left[\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\mathbf{y}\right)}\right] g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep} \\ = & \int g\left(\mathbf{y}_{rep}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep} - \int p\left(\mathbf{y}_{rep}|\mathbf{y}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep}, \end{aligned} \tag{3}$$

where $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ is a predictive distribution. Note that the first term is the same across all candidate models which is denoted by C. Thus, we get

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

If one chooses $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})\right)$, where $\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})$ is the MLE of $\boldsymbol{\theta}$ based on $\mathbf{y}$, then it is well-known that (see, for example, Burnham and Anderson (2002))

$$
\begin{aligned}
& E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})\right)\right]\right\} = 2C + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})\right)\right] \\
= \; & 2C + E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})\right) + 2P\right) + o(1) = 2C + E_{\mathbf{y}}\left(\text{AIC}\right) + o(1),
\end{aligned}
$$

where the expectation $E_{\mathbf{y}}$ and $E_{\mathbf{y}_{rep}}$ are related to $g\left(\mathbf{y}\right)$ and $g\left(\mathbf{y}_{rep}\right)$, respectively. Hence, AIC is an asymptotically unbiased estimator of the expected KL divergence minus a constant, that is, $E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_{ML}(\mathbf{y})\right)\right]\right\} - 2C$.

If one chooses $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$, where $\bar{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}$ based on $\mathbf{y}$, Li, et al (2017) showed that

$$
\begin{aligned}
& E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]\right\} = 2C + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] \\
= \; & 2C + E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right) + 2P_D\right) + o(1) = 2C + E_{\mathbf{y}}\left(\text{DIC}\right) + o(1). \qquad (4)
\end{aligned}
$$

Hence, DIC is an asymptotically unbiased estimator of the expected KL divergence minus a constant, that is, $E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]\right\} - 2C$.

The smaller AIC/DIC, the better predictive performance of the candidate model. When the prior information is dominated by likelihood, Li, et al (2017) also showed that DIC and AIC are asymptotically equivalent, i.e.,

$$
\text{DIC} = \text{AIC} + o_p(1), P_D = P + o_p(1).
$$

This explains why DIC has been explained as a Bayesian version of AIC in the literature.

As pointed out by Spiegelhalter, et al (2014), the plug-in predictive distribution for DIC is not a proper predictive distribution and not invariant to reparametrization. Based on the Bayesian predictive distribution, Li, et al. (2017) proposed the following version of DIC (named DIC^{BP}) for Bayesian model comparison,

$$
\text{DIC}^{BP} = D(\bar{\boldsymbol{\theta}}) + (1 + \ln 2)P_D, \qquad (5)
$$

When choosing $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) to be the full Bayesian predictive distribution $p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$, Li, et al. (2017) showed that

$$
\begin{aligned}
& E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right]\right\} = 2C + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[-2\ln p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] \\
= \; & 2C + E_{\mathbf{y}}\left(\text{DIC}^{BP}\right) + o(1), \qquad (6)
\end{aligned}
$$

Hence, DIC^{BP} is an asymptotically unbiased estimator of the expected KL divergence which measures the quality of the candidate model in terms of its ability to make the full Bayesian prediction, minus a constant. Hence, the smaller DIC^{BP}, the better predictive performance of the candidate model using the Bayesian predictive distribution. Since $p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ is invariant to reparametrization, DIC^{BP} is also invariant to reparametrization.

When deriving the asymptotic theory given in (4) and (6), Li, et al (2017) had to impose a set of regularity conditions. Essentially these conditions ensure the following key asymptotic properties. First, the Bernstein-von Mises theorem holds. That is, the posterior distribution converges to a normal distribution with the MLE as its mean and the inverse of the second derivative of the negative log-likelihood function evaluated at the MLE as its covariance. Second, the standard large sample theory for ML holds, including consistency, asymptotic normality with the covariance being the inverse of the second derivative of the negative log-likelihood function evaluated at the true parameter value. Third, the difference between the posterior mean and the MLE is $O_p(n^{-1})$. Fourth, the difference between the posterior covariance and the asymptotic covariance of MLE is $O_p(n^{-2})$.

3 DIC for Latent Variable Models

3.1 MCMC and data augmentation

Let $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ denote observed data and $\mathbf{z} = (z_1, z_2, \cdots, z_n)'$ be latent variables.[1] Let a latent variable model be indexed by the a set of P parameters, $\boldsymbol{\theta} = (\theta_1, \dots, \theta_P)' \in \boldsymbol{\Theta} \subseteq R^P$. 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:

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta}) d\mathbf{z}. \tag{7}$$

Typically the integral in (7) does not have a closed-form solution. Consequently, the ML method is difficult to use as it requires calculations of $p(\mathbf{y}|\boldsymbol{\theta})$ for each value of $\boldsymbol{\theta}$ during numerical optimizations.

If posterior analysis is conducted based on the observed-data likelihood, $p(\mathbf{y}|\boldsymbol{\theta})$, one would end up with the same problem as in ML since $p(\mathbf{y}|\boldsymbol{\theta})$ does not have a closed-form expression. An alternative way to conduct posterior analysis is to treat $\mathbf{z}$ as parameters. Consequently, the new likelihood function becomes $p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ (i.e. conditional likelihood) which is often available in closed-form. In the Bayesian literature, this parameter expansion technique is known as data augmentation. The closed-form expression in the new likelihood function facilitates MCMC sampling from the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$.

After a sufficiently long period for a burn-in phase, the simulated random samples can be regarded as random observations from the joint distribution. The statistical analysis can be established from 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

[1] Although we assume that the number of latent variables is the same as that of the observed data points, such an assumption may be relaxed. A more general assumption is that the number of latent variables grows proportionally with that of the observed data points. In this more general case, the theory discussed below continues to hold.

posterior analysis 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 the Bayesian analysis of latent variable models which is a powerful alternative to ML.

3.2 DIC for latent variable models

As described in Section 3.1, in a latent variable model, 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 function, $p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})\mathrm{d}\mathbf{z}$, is clearly defined. In this case, only $\boldsymbol{\theta}$, not $\mathbf{z}$, is treated as parameters. In the Bayesian framework, however, depending on whether the latent variables $\mathbf{z}$ are treated as parameters or not, three likelihood functions may be used, $p(\mathbf{y}|\boldsymbol{\theta})$, $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$, and $p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ which correspond to observed-data likelihood, complete-data likelihood and conditional likelihood.

With these three likelihood functions, Celeux et al (2006) considered and compared eight versions of DIC. Based on $p(\mathbf{y}|\boldsymbol{\theta})$, the first three versions are

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

where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mode. It is easy to show that DIC_1 is the same as DIC given in (1). Based on $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$, next three versions are

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

where in DIC_4, $\bar{\boldsymbol{\theta}}\left(\mathbf{y}, \mathbf{z}\right)$ is the posterior mean estimator of $\boldsymbol{\theta}$ based on $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$; in DIC_5, $\hat{\mathbf{z}}_{JE}(\mathbf{y})$ and $\hat{\boldsymbol{\theta}}_{JE}(\mathbf{y})$ are the joint estimator, such as the posterior mean or the posterior mode of $(\mathbf{z}, \boldsymbol{\theta})$. Based on $p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$, the last two versions are

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

where in DIC_8, $\hat{\boldsymbol{\theta}}(\mathbf{y}, \mathbf{z})$ is an estimator of $\boldsymbol{\theta}$ based on $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$, such as posterior mean or the posterior mode.

To determine which likelihood is used for constructing DIC, Spiegelhalter, et al (2002) and Celeux et al (2006) both used a notion called "focus". If only $\boldsymbol{\theta}$ 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 DIC_1 and DIC_2. If both $\mathbf{z}$ and $\boldsymbol{\theta}$ are in "focus", the conditional likelihood $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ is used for constructing DIC. This choice of "focus" leads to DIC_7 and DIC_8. Clearly, the other three choices, namely DIC_4, DIC_5 and 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, DIC_8 is logically incoherent because parameters in "focus" are $(\mathbf{z},\boldsymbol{\theta})$ in the first term, but become $\mathbf{z}$ in the second term. As pointed out by Plummer (2006), DIC_3 does not have a "focus" corresponding to it and it is not clear which likelihood is used to construct DIC_3. Therefore, only DIC_1, DIC_2, and DIC_7 are logically coherent. Furthermore, Celeux et al (2006) compared DIC_1 with DIC_2 and found the evidence that DIC_2 is better than DIC_1 in the sense that the posterior mode, but not the posterior mean, ensures positivity of P_D. However, under the set of regularity conditions listed below, we can show that DIC_1 and DIC_2 are asymptotically equivalent. In practice, the posterior mode is more difficult to compute than the posterior mean. Hence, from a computational viewpoint, it is easier to obtain DIC_1 than DIC_2, making DIC_1 more popular in practice.

Given the discussion above, not surprisingly, DIC_1 is monitored and reported in WinBUGS when there is no latent variable. To compute DIC_1, it is generally required that observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ be available in closed-form because $E_{\boldsymbol{\theta}|\mathbf{y}}\left[\ln p(\mathbf{y}|\boldsymbol{\theta}))\right]$ may be arbitrarily well approximated by $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ for a large J and $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ is easy to compute. The observed likelihood function is often available in closed-form when there is no latent variable.

Unfortunately, for many latent variable models, such as state-space models, the observed-data likelihood $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form. In this case, DIC_1 is difficult to compute because it needs to evaluate the observed-data likelihood for J times. Given that J is usually large, computing $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ without an analytical expression for $\ln p(\mathbf{y}|\boldsymbol{\theta}))$ is time consuming. In DIC_7, the latent variables are regarded as parameters and $\ln p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ often has an analytical expression. Hence, it is easy to compute $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\mathbf{z}^{(j)},\boldsymbol{\theta}^{(j)}\right)$. That is why, when there are latent variables, data augmentation is used to obtain Markov chains for both $\mathbf{z}$ and $\boldsymbol{\theta}$. Based on the MCMC output for $\mathbf{z}$ and $\boldsymbol{\theta}$ and by choosing the deviance based on $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$, DIC_7 can be easily computed. Following the suggestion of Spiegelhalter, et al (2002), DIC_7 is monitored and reported in WinBUGS for latent variable models. Clearly the use of DIC_7 is for computational convenience, as explained in Spiegelhalter, et al (2002). The corresponding "focus" contains both $\mathbf{z}$ and $\boldsymbol{\theta}$ due to data augmentation.

However, from a theoretical viewpoint, DIC_7 has a few problems. Firstly, with data augmentation, the dimension of the parameter space is much bigger, increasing from P to $n+P$. Since the dimension of the parameter space grows proportionally with the number of data points, the new likelihood function $p(\mathbf{y}|\mathbf{z},\boldsymbol{\theta})$ is not regular and it leads to the well-known incidental parameter problem in econometrics 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. For example, the ML estimator of $\mathbf{z}$ is inconsistent as its variance does not go to zero as $n \to \infty$. Similarly, the Bayesian large sample theory becomes invalid; see Page 89-90 of Gelman, et al (2013). Obviously, the failure of the standard asymptotic theory invalidates the asymptotic justification of DIC$_7$. In fact, it also invalidates the asymptotic justification of AIC if AIC is constructed from conditional likelihood.

To illustrate this problem, 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 MLE of σ^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 normally distributed. However, if $\{\alpha_i\}_{i=1}^n$ are treated as parameters, they are incidental in the sense of Neyman and Scott (1948). The MLE of α_i is $\widehat{\alpha}_i = y_i \sim N(\alpha_i, \sigma^2)$ which is correctly centered at α_i but inconsistent as the variance of MLE does not go to zero as n grows. If $\sigma^2 = 1$ and is assumed to be known, then $P = n$ and the posterior distribution is $\alpha_i|y_i \sim N(0.5y_i, 0.5)$. The posterior mean (which is also the posterior mode) is $\overline{\alpha}_i = 0.5y_i$ which is not centered at the MLE. The posterior variance is 0.5 which does not go to zero as n grows. Clearly, both the standard ML large sample theory and the Bayesian large sample theory fail to hold. These results are not surprising as only one observation (y_i) contains information about α_i.

Let $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_n)'$ and $\tilde{\boldsymbol{\alpha}}(\mathbf{y})$ be an estimator of $\boldsymbol{\alpha}$. By evaluating (3) we have

$$KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\alpha}}(\mathbf{y})\right)\right] = E_{\mathbf{y}_{rep}}\left[\ln\frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\tilde{\boldsymbol{\alpha}}\left(\mathbf{y}\right)\right)}\right]$$

$$\begin{aligned} &= C - \int \ln p\left(\mathbf{y}^{rep}|\tilde{\boldsymbol{\alpha}}(\mathbf{y})\right) g(\mathbf{y}^{rep}) \mathrm{d}\mathbf{y}^{rep} \\ &= C + \left[\frac{n}{2}\ln(2\pi\sigma^2) + \frac{n\left(\sigma^2+1\right)}{2\sigma^2} + \sum_{i=1}^n \frac{\tilde{\alpha}_i^2\left(\mathbf{y}\right)}{2\sigma^2}\right]. \end{aligned} \tag{8}$$

When $\sigma^2 = 1$, by plugging the MLE of α_i (i.e. $\widehat{\alpha}_i = y_i$) into (8), multiplying both sides by 2 and taking expectation with respect to $\mathbf{y}$, we have

$$\begin{aligned} E_{\mathbf{y}}\left(2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n\right)\right] - 2C\right) &= n\ln(2\pi) + 2n + \sum_{i=1}^n E\left(y_i^2\right) \\ &= n\ln(2\pi) + 4n. \end{aligned}$$

However,

$$E_{\mathbf{y}}(\mathrm{AIC}) = E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\widehat{\alpha}_1, \ldots, \widehat{\alpha}_n\right)\right) + 2n = n\ln(2\pi) + 2n.$$

Similarly, by plugging the posterior mean of α_i (i.e. $\overline{\alpha}_i = 0.5y_i$) into (8), multiplying both sides by 2 and taking expectation with respect to $\mathbf{y}$, we have

$$\begin{aligned} E_{\mathbf{y}}\left(2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\overline{\alpha}_1, \dots, \overline{\alpha}_n\right)\right] - 2C\right) &= n\ln(2\pi) + 2n + \sum_{i=1}^{n} \frac{E\left(y_i^2\right)}{4} \\ &= n\ln(2\pi) + 2.5n. \end{aligned}$$

However,

$$\begin{aligned} P_D &= -2\int \left[\ln p(\mathbf{y}|\boldsymbol{\alpha}) - \ln p\left(\mathbf{y}|\overline{\alpha}_1, \dots, \overline{\alpha}_n\right)\right] p(\boldsymbol{\alpha}|\mathbf{y})\mathrm{d}\boldsymbol{\alpha} \\ &= -2\int \left[\ln p(\mathbf{y}|\boldsymbol{\alpha})\right] p(\boldsymbol{\alpha}|\mathbf{y})\mathrm{d}\boldsymbol{\alpha} + 2\ln p\left(\mathbf{y}|\overline{\alpha}_1, \dots, \overline{\alpha}_n\right) \\ &= \sum_{i=1}^{n} \int (y_i - \alpha_i)^2 p(\alpha_i|y_i)d\alpha_i - \frac{\sum_{i=1}^{n} y_i^2}{2} \\ &= \sum_{i=1}^{n}\left[\frac{1}{2} + \frac{y_i^2}{4}\right] - \frac{\sum_{i=1}^{n} y_i^2}{2} = \frac{n}{2} - \frac{\sum_{i=1}^{n} y_i^2}{4}, \end{aligned}$$

$$\begin{aligned} E_{\mathbf{y}}\left(\mathrm{DIC}\right) &= E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\overline{\alpha}_1, \dots, \overline{\alpha}_n\right) + 2P_D\right) \\ &= E_{\mathbf{y}}\left(n\ln(2\pi) + \frac{\sum_{i=1}^{n} y_i^2}{2} + 2P_D\right) \\ &= n\ln(2\pi) + n. \end{aligned}$$

Thus,

$$\begin{aligned} E_{\mathbf{y}}\left(2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\widehat{\alpha}_1, \dots, \widehat{\alpha}_n\right)\right] - 2C\right) &= E_{\mathbf{y}}(\mathrm{AIC}) + 2n, & (9) \\ E_{\mathbf{y}}\left(2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\overline{\alpha}_1, \dots, \overline{\alpha}_n\right)\right] - 2C\right) &= E_{\mathbf{y}}(\mathrm{DIC}) + 1.5n, & (10) \\ E_{\mathbf{y}}(P_D) &= 0 \neq n + o(1), & (11) \\ E_{\mathbf{y}}(\mathrm{AIC} - \mathrm{DIC}) &= n \neq o_p(1). & (12) \end{aligned}$$

According to (9), AIC is a biased estimator of the corresponding expected KL divergence minus a constant asymptotically. According to (10), DIC is a biased estimator of the corresponding expected KL divergence minus a constant asymptotically. According to (11), on average the effective number of parameter (P_D) is zero. According to (12), on average AIC differs from DIC by n. All these observations are at odds with the theory discussed earlier.

Secondly, sometimes a statistical model without latent variable can be represented by another model with latent variables. A leading example in the Bayesian literature is the Student t distribution which can be rewritten as a normal-inverse-gamma distribution where the variance is assumed to follow an inverse-gamma distribution and hence is treated as a latent variable. These two equivalent representations, even under the same priors, often lead to very different DIC values. The reason for this sharp discrepancy is that in the model without latent variables, DIC_1 is used while in the model with latent variables, DIC_7 is used.

This problem arises in Section 8.2 of Spiegelhalter, et al (2002) and in Model 8 of Berg, et al (2004).

Thirdly, due to data augmentation, the dimension of the parameter space becomes much larger and hence DIC_7 is expected to be sensitive to transformations of latent variables. To illustrate this 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 both models there are latent variables. In Model 2 the latent variable is the volatility σ_t^2 while the latent variable is the log-volatility $h_t = \ln \sigma_t^2$ in Model 1. Hence, following the usual practice in the literature, DIC_7 is the relevant version. Since the two models are identical, we expect the two models give the same DIC_7 value. To calculate 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 data augmentation, the latent variables, h_t and σ_t^2 are regarded as parameters, and we find that $P_D = 89.806$ and $\text{DIC}_7 = 2884.37$ for Model 1 but $P_D = 59.366$ and $\text{DIC}_7 = 2852.85$ 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 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 DIC_7 is easier to calculate and has been used widely in practice but suffers from several theoretical problems, While DIC_1 has rigorously theoretical justification, it is very hard to compute from MCMC output since $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form.

There are several recent studies that document the problem with DIC_7. In the context of negative binomial and Poisson-lognormal models, Millar (2009) found strong evidence of poor performance of DIC_7 using simulated data. For example, he simulated data from a Poisson-lognormal model but found that DIC_7 almost always prefers the Poisson-gamma model instead of the Poisson-lognormal model. Millar and McKechnie (2014) documented the same problem of DIC_7 in state-space models using simulated data. To deal with the problem, they suggested a one-step-ahead DIC, where prediction is conditional on the state at the previous time point. Chan and Grant (2016a) showed that, in the context of stochastic volatility models, DIC_7 tends to favor overfitted models in a Monte Carlo study. To deal with the problem, they suggested using DIC$_1$. To compute DIC$_1$, they introduced an important-sampling-based algorithm. In the context of three classes of latent variable models (namely factor models, linear Gaussian state-space models and semiparametric regression models), Chan and Grant (2016b) developed fast algorithms based on sparse matrix algorithms to compute DIC$_1$. The proposed algorithms require repeated evaluations of observed-data likelihood. For models where observed-data likelihood cannot be quickly evaluated, such as general nonlinear random-Gaussian models, it is difficult to calculate DIC$_1$.

4 Integrated DIC for Latent Variable Models

Based on the discussion above, DIC$_7$ lacks of theoretical justification and DIC$_1$ is difficult to compute for latent variable models. There is a great need to introduce a model selection criterion which has theoretical justification, and is generally applicable to general latent variable models and feasible to compute. In this section, we propose two versions of DIC (denoted as integrated-likelihood DIC or IDIC) based on two different predictive distributions.

4.1 IDIC based on plug-in predictive distribution

When $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ in (3) is chosen to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$, where $\bar{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}$ on the data $\mathbf{y}$ (when there is no confusion, we simple write $\bar{\boldsymbol{\theta}}(\mathbf{y})$ as $\bar{\boldsymbol{\theta}}$), we propose the following IDIC,

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

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

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

and

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

Clearly, the leading term in IDIC is the same as that in DIC$_1$. However, the penalty term in DIC$_1$ is $2P_D$ while it is $2P_D^I$ in IDIC.

To theoretically justify IDIC, we will develop the large sample properties of IDIC under some regularity conditions, that is, in the same spirit as how DIC$_1$ has been justified by Li, et al (2017). In particular, we will show that IDIC can approximate AIC and P_D^I can approximate P, the number of parameters. The order for approximation errors is given. Consequently, IDIC provides asymptotically unbiased estimation to the KL divergence based on the plug-in predictive distribution.

Let $\mathbf{y}$ be a collection of random variables defined on a probability space $\{\Omega, \mathcal{F}, \wp_\theta\}$, where $\wp_\theta$ is a probability measure that depends on parameter $\boldsymbol{\theta} \in \Theta$, which is a compact subset of R^P. Let $\mathbf{y}^t := (y_0, y_1, \ldots, y_t)$ for any $0 \leq t \leq n$ and $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right) = \ln p(\mathbf{y}^t|\boldsymbol{\theta}) - \ln p(\mathbf{y}^{t-1}|\boldsymbol{\theta})$ be the conditional log-likelihood for the t^{th} observation for any $1 \leq t \leq n$. When there is no confusion, we suppress $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right)$ as $l_t(\boldsymbol{\theta})$ so that the log-likelihood function $\ln p(\mathbf{y}|\boldsymbol{\theta})$ is $\sum_{t=1}^n l_t(\boldsymbol{\theta})$.[2] And define $l_t^{(j)}(\boldsymbol{\theta})$ to be the j^{th} derivative of $l_t(\boldsymbol{\theta})$ and $l_t^{(j)}(\boldsymbol{\theta}) = l_t(\boldsymbol{\theta})$ when $j = 0$. We introduce the following functions

$$\mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) := \frac{\partial \ln p(\mathbf{y}^t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^t l_i^{(1)}(\boldsymbol{\theta}),\ \mathbf{H}(\mathbf{y}^t, \boldsymbol{\theta}) := \frac{\partial^2 \ln p(\mathbf{y}^t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} = \sum_{i=1}^t l_i^{(2)}(\boldsymbol{\theta}),$$

$$\mathbf{s}_t(\boldsymbol{\theta}) := l_t^{(1)}(\boldsymbol{\theta}) = \mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{s}(\mathbf{y}^{t-1}, \boldsymbol{\theta}),\ \mathbf{H}_t(\boldsymbol{\theta}) := l_t^{(2)}(\boldsymbol{\theta}) = \mathbf{H}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{H}(\mathbf{y}^{t-1}, \boldsymbol{\theta}),$$

$$\mathbf{B}_n(\boldsymbol{\theta}) := Var\left[\frac{1}{\sqrt{n}} \sum_{t=1}^n l_t^{(1)}(\boldsymbol{\theta})\right], \bar{\mathbf{H}}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{t=1}^n \mathbf{H}_t(\boldsymbol{\theta}),$$

$$\bar{\mathbf{J}}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{t=1}^n \left[\mathbf{s}_t(\boldsymbol{\theta}) - \bar{\mathbf{s}}_n(\boldsymbol{\theta})\right] \left[\mathbf{s}_t(\boldsymbol{\theta}) - \bar{\mathbf{s}}_n(\boldsymbol{\theta})\right]', \bar{\mathbf{s}}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^n \mathbf{s}_t(\boldsymbol{\theta}),$$

$$\mathbf{H}_n(\boldsymbol{\theta}) := \int \bar{\mathbf{H}}_n(\boldsymbol{\theta}) g(\mathbf{y})\, d\mathbf{y},\ \mathbf{J}_n(\boldsymbol{\theta}) = \int \bar{\mathbf{J}}_n(\boldsymbol{\theta}) g(\mathbf{y})\, d\mathbf{y},$$

In this paper, as in Li, et al (2017), we impose the following regularity conditions.

Assumption 1: $\boldsymbol{\Theta} \subset R^P$ is compact.

Assumption 2: $\{y_t\}_{t=1}^{\infty}$ satisfies the strong mixing condition with the mixing coefficient $\alpha(m) = O\left(m^{\frac{-2r}{r-2} - \varepsilon}\right)$ for some $\varepsilon > 0$ and $r > 2$.

Assumption 3: For all t, $l_t(\boldsymbol{\theta})$ satisfies the standard measurability and continuity condition, and the eight-times differentiability condition on $F_{-\infty}^t \times \boldsymbol{\Theta}$ where $F_{-\infty}^t = \sigma(y_t, y_{t-1}, \cdots)$**.**

Assumption 4: For $j = 0, 1, 2$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\left\| l_t^{(j)}(\boldsymbol{\theta}) - l_t^{(j)}(\boldsymbol{\theta}') \right\| \leq c_t^j\left(\mathbf{y}^t\right) \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$ in probability, where $c_t^j\left(\mathbf{y}^t\right)$ is a positive random variable with $\sup_t E\left\| c_t^j\left(\mathbf{y}^t\right) \right\| < \infty$ and $\frac{1}{n}\sum_{t=1}^n \left(c_t^j\left(\mathbf{y}^t\right) - E\left(c_t^j\left(\mathbf{y}^t\right)\right)\right) \xrightarrow{p} 0$.

Assumption 5: For $j = 0, 1, \ldots, 8$, there exists a function $M_t(\mathbf{y}^t)$ such that for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, $l_t^{(j)}(\boldsymbol{\theta})$ exists, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| l_t^{(j)}(\boldsymbol{\theta}) \right\| \leqslant M_t(\mathbf{y}^t)$, and $\sup_t E\left\| M_t(\mathbf{y}^t) \right\|^{r+\delta} \leq M < \infty$ for some $\delta > 0$, where r is the same as that in Assumption 2.

Assumption 6: $\left\{ l_t^{(j)}(\boldsymbol{\theta}) \right\}$ is L_2-near epoch dependent with respect to $\{\mathbf{y}_t\}$ of size -1 for $0 \leqslant j \leqslant 1$ and $-\frac{1}{2}$ for $j = 2$ uniformly on $\boldsymbol{\Theta}$.

Assumption 7: Let $\boldsymbol{\theta}_n^p$ be the pseudo-true value that minimizes the KL loss between the DGP and the candidate model

$$\boldsymbol{\theta}_n^p = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \frac{1}{n} \int \ln \frac{g(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})} g(\mathbf{y}) d\mathbf{y},$$

where $\{\boldsymbol{\theta}_n^p\}$ is the sequence of minimizers interior to $\boldsymbol{\Theta}$ uniformly in n. For all $\varepsilon > 0$,

$$\lim_{n \to \infty} \sup \sup_{\Theta \backslash N(\boldsymbol{\theta}_n^p, \varepsilon)} \frac{1}{n} \sum_{t=1}^n \left\{ E\left[l_t(\boldsymbol{\theta})\right] - E\left[l_t(\boldsymbol{\theta}_n^p)\right] \right\} < 0, \tag{17}$$

[2] In the definition of log-likelihood, we ignore the initial condition $\ln p(y_0)$. For weakly dependent data, the impact of the initial condition is asymptotically negligible.

where $N\left(\boldsymbol{\theta}_n^p, \varepsilon\right)$ is the open ball of radius ε around $\boldsymbol{\theta}_n^p$.

Assumption 8: The sequence $\left\{\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right)\right\}$ is negative definite and the sequence $\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\right\}$ is positive definite, both uniformly in n.

Assumption 9: $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) = o\left(1\right)$.

Assumption 10: The prior density $p(\boldsymbol{\theta})$ is eight-times continuously differentiable, $p(\boldsymbol{\theta}_n^p) > 0$ and $\int \left\|\boldsymbol{\theta}\right\|^2 p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta} < \infty$.

Lemma 4.1 below gives a high order approximation to the posterior mean and the posterior variance based on a high order Laplace expansion. To apply the Laplace expansion, we need to fix more notations. For convenience of exposition, we let $\overline{\mathbf{H}}_n^{(j)}\left(\boldsymbol{\theta}\right) = \frac{1}{n}\sum_{t=1}^n l_t^{(j)}\left(\boldsymbol{\theta}\right)$ for $j = 3, 4, 5$. Let $\pi\left(\boldsymbol{\theta}\right) = \ln p\left(\boldsymbol{\theta}\right)$, $p^{(j)}\left(\boldsymbol{\theta}\right)$, $\pi^{(j)}\left(\boldsymbol{\theta}\right)$ be the jth order derivatives of $p\left(\boldsymbol{\theta}\right)$, $\pi\left(\boldsymbol{\theta}\right)$ for $j = 1, 2$, and $\widehat{p}$, $\widehat{\pi}$, $\widehat{p}^{(j)}$ and $\widehat{\pi}^{(j)}$ be the values of functions $p\left(\boldsymbol{\theta}\right)$, $\pi\left(\boldsymbol{\theta}\right)$, $p^{(j)}\left(\boldsymbol{\theta}\right)$ and $\pi^{(j)}\left(\boldsymbol{\theta}\right)$ evaluated at $\widehat{\boldsymbol{\theta}}_{ML}(\mathbf{y})$. When there is no confusion, we simply write $\widehat{\boldsymbol{\theta}}_{ML}(\mathbf{y})$ as $\widehat{\boldsymbol{\theta}}$**.**

Lemma 4.1 *Let $Var(\boldsymbol{\theta}|\mathbf{y}) = E\left[(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})(\boldsymbol{\theta} - \bar{\boldsymbol{\theta}})'|\mathbf{y}\right]$ be the posterior variance of $\boldsymbol{\theta}$. Under Assumptions 1-10, it can be shown that*

$$\bar{\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right),$$

$$vec\left[Var(\boldsymbol{\theta}|\mathbf{y})\right] = -\frac{1}{n}vec\left(\bar{\mathbf{H}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}\right)\right) + \frac{1}{n^2}(F_1 + F_2) + O_p\left(\frac{1}{n^3}\right),$$

where

$$B_1^1 = \bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} - \frac{1}{2}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right),$$

$$\begin{aligned}
B_2^1 \;=\; & -\frac{1}{8}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(5)\prime}vec\left[\bar{\mathbf{H}}_n^{-1}\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right] \\
& +\frac{35}{48}\left[\bar{\mathbf{H}}_n^{-1}\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right]'\bar{\mathbf{H}}_n^{(4)}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right) \\
& -\frac{35}{48}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\left[vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right] \\
& -\frac{5}{8}\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right)\bar{\mathbf{H}}_n^{(4)\prime}\right] \\
& +\frac{35}{24}\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}\left[vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right] \\
& -\frac{5}{4}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] + \frac{1}{2}\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(3)}}{\widehat{p}}'vec\left(\bar{\mathbf{H}}_n^{-1}\right),
\end{aligned}$$

$$B_3^1 = B_1^1 \times B_4^1,$$

$$
\begin{aligned}
B_4^1 &= \frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] - \frac{1}{2}vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} \\
&\quad + \frac{5}{24}\left[vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right] \\
&\quad + \frac{1}{8}\mathbf{tr}\left[\left[\bar{\mathbf{H}}_n^{-1}\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right]\bar{\mathbf{H}}_n^{(4)\prime}\right],
\end{aligned}
$$

$$
\begin{aligned}
F_1 &= -\frac{7}{16}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n^{-1}\otimes vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\right] \\
&\quad + \frac{25}{48}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\left[vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\right],
\end{aligned}
$$

$$
\begin{aligned}
F_2 &= -\frac{5}{2}vec\left[\bar{\mathbf{H}}_n^{-1}\bar{\mathbf{H}}_n^{(3)\prime}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\frac{\widehat{p}^{(1)}{}'}{\widehat{p}}\bar{\mathbf{H}}_n^{-1}\right] + \frac{1}{4}vec\left(\bar{\mathbf{H}}_n^{-1}\right)\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] \\
&\quad + \frac{1}{2}vec\left(\bar{\mathbf{H}}_n^{-1}\right)vec\left(\bar{\mathbf{H}}_n^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\bar{\mathbf{H}}_n^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}},
\end{aligned}
$$

vec denotes the column-wise vectorization of a matrix, and **tr** *denotes the trace of a matrix.*

Remark 4.1 *Under the different regularity conditions, the Bernstein-von Mises theorem shows that the posterior distribution converges to a normal distribution with the MLE as its mean and the inverse of the second derivative of the negative log-likelihood function evaluated at the MLE as its variance. Based on the Bernstein-von Mises theorem, when the parameter is one-dimension, Ghosh and Ramamoorthi (2003) developed the similar results with Lemma 4.1 for the iid case. In particular, Ghosh and Ramamoorthi (2003) showed that*

$$
\bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} = o_p(n^{-1/2}), Var(\boldsymbol{\theta}|\mathbf{y}) + \frac{1}{n}\bar{\mathbf{H}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}\right) = o_p(n^{-1}).
$$

Our Lemma 4.1 extend the results of Ghosh and Ramamoorthi (2003) in three aspects: (1) to the weakly dependent case; (2) to the multivariate-dimension case; (3) giving the exact order of the first and second moments of the difference between the posterior distribution and the asymptotic normal distribution. From Lemma 4.1, we can easily obtain that

$$
\bar{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}} = O_p(n^{-1}), Var(\boldsymbol{\theta}|\mathbf{y}) + \frac{1}{n}\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right) = O_p(n^{-2}).
$$

Based on this lemma, we can obtain the exact order of the difference between IDIC and AIC as follows.

Theorem 4.1 *Under Assumptions 1-10, we have*

$$
P_D^I = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(\frac{1}{n^2}\right),
$$

$$IDIC = AIC + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(\frac{1}{n^2}\right),$$

where

$$C_1 = \frac{7P}{16}C_{11} + \frac{24-25P}{48}C_{12}, C_2 = \frac{3-P}{2}C_{21} - \frac{P}{4}C_{22} - \frac{P}{4}C_{23},$$

$$D_1 = \frac{7P}{8}C_{11} + \frac{18-25P}{24}C_{12}, D_2 = (4-P)\,C_{21} - \frac{P}{2}C_{22} - \frac{2+P}{2}C_{23},$$

$$C_{11} = \mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1} \otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right],$$

$$C_{12} = vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right),$$

$$C_{21} = vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right),$$

$$C_{22} = \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\right], C_{23} = \pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right)'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right).$$

Corollary 4.2 *Under Assumptions 1-10, we have*

$$\begin{aligned} & E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}\right)\right]\right\} = 2C + E_{\mathbf{y}}\left[-2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) + 2P_D^I\right] + o(1) \\ = \quad & 2C + E_{\mathbf{y}}\left(IDIC\right) + o(1). \end{aligned}$$

Remark 4.2 *In Equation (15) on Page 590, Spiegelhalter, et al. (2002) obtained the expression for P_D^I and claimed that P_D^I approximates the P_D component in DIC_1 and P in AIC. Unfortunately, to the best of our knowledge, P_D^I has never been implemented in practice and WinBUGS does not report P_D^I. Moreover, the conditions under which $P_D^I \approx P_D \approx P$ holds true were not specified in Spiegelhalter, et al (2002). The order of the approximation error was unknown. According to Theorem 4.1, the order of the difference between P and P_D^I and that between AIC and IDIC are both $O_p(n^{-1})$. Furthermore, combined with Lemma 3.3 in Li et al (2017), it is easy to show that the order of approximation error between P_D and P_D^I and that between DIC_1 and IDIC are also $O_p(n^{-1})$.*

Remark 4.3 *Theorem 4.1 clearly shows that the order of difference between AIC and IDIC is $O_p(n^{-1})$. For this reason, both IDIC and DIC_1 can be regarded as the Bayesian version of AIC. When the prior is informative and the sample size n is finite, IDIC may give different value from AIC. Like DIC_1, an important contribution of IDIC is that it provides an approach to measure the model complexity when the informative prior is available.*

Remark 4.4 *Corollary 4.2 is the direct result of Theorem 4.1 and Theorem 3.1 of Li, et al (2017). It gives the decision-theoretical justification of IDIC. As DIC$_1$, IDIC is also an asymptotically unbiased estimator of the expected KL divergence minus a constant, that is, $E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}\right)\right]\right\}-2C$. Hence, as DIC$_1$, IDIC selects a model that minimizes the expected KL divergence between the DGP and the plug-in predictive distribution. The smaller the value of IDIC, the better the predictive performance of the candidate latent variable model.*

Remark 4.5 *From the discussion above, DIC$_1$ and IDIC share the same asymptotic properties. However, as explained before, there is an important difference between DIC$_1$ and IDIC, that is, the penalty term takes a different expression. It is this difference that makes IDIC easier to compute from MCMC output. To compute P_D in DIC$_1$, one has to evaluate $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ and hence calculate $p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ for J times. For latent variable models, since $p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ is not available in closed-form, the computational cost is high. However, to compute P_D^I in IDIC, one needs to evaluate the second derivative of observed-data likelihood only once, which is computationally much less expensive. In Section 4.3, we will introduce some efficient algorithms to evaluate $D(\bar{\boldsymbol{\theta}})$ and $\mathbf{I}(\bar{\boldsymbol{\theta}})$.*

Remark 4.6 *In the context of latent variable models, while DIC$_7$ is trivial to calculate but cannot be theoretically justified, DIC$_1$ is theoretically justified but hard to compute. IDIC solves this dilemma because it is theoretically justified and computational inexpensive. The corresponding deviance is based on the observed-data likelihood function and the latent variables are not treated as parameter. It is important to point out that IDIC is computed from MCMC output. While IDIC does not treat latent variables as parameters, MCMC output may be obtained based on the data augmentation technique without affecting the asymptotic justification of IDIC. Return to the Clark model, with the same setting as before, we get $P_D^I = 1.75$ for Model 1 and $P_D^I = 1.80$ 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 in the model. This is what we expect given that the vague priors are used. The small difference between P_D^I and P arises due to the simulation error and the priors.*

4.2 IDIC based on Bayesian predictive distribution

It is well-known that DIC is not invariant to reparametrization; see Spiegelhalter, et al (2014). This problem motivated Li, et al (2017) to introduce DIC^{BP} based on the Bayesian predictive distribution. Li, et al (2017) shows that DIC^{BP} is asymptotically unbiased for estimating the expected loss function associated with the KL divergence between the true DGP and the Bayesian predictive distribution minus a constant. For models without latent variables, DIC^{BP} of Li, et al takes the form

$$\text{DIC}^{BP} = D(\bar{\boldsymbol{\theta}}) + (1 + \ln 2)P_D.$$

For latent variable models, just as DIC$_1$, DIC^{BP} is difficult to compute. Hence, we propose a version of IDIC based on the Bayesian predictive distribution (call it as IDIC^{BP}) which is defined as

$$\text{IDIC}^{BP} = D(\bar{\boldsymbol{\theta}}) + (1 + \ln 2)P_D^I = D(\bar{\boldsymbol{\theta}}) + (1 + \ln 2)\mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}. \tag{18}$$

Theorem 4.3 *Under Assumptions 1-10, it can be shown that*

$$\begin{aligned} & IDIC^{BP} = DIC^{BP} + o_p(1), \\ & E_{\mathbf{y}}\left\{2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right]\right\} = 2C + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] \\ = \quad & 2C + E_{\mathbf{y}}\left(DIC^{BP}\right) + o(1) = 2C + E_{\mathbf{y}}\left(IDIC^{BP}\right) + o(1). \end{aligned}$$

Remark 4.7 *Theorem 4.3 is the direct result of Theorem 4.1 and Theorem 4.1 of Li, et al (2017). According to this corollary, IDIC^{BP} is an asymptotically unbiased estimator of the expected KL divergence which measures the quality of the candidate model in terms of its ability to make predictions using the Bayesian predictive distribution. Hence, the smaller the value of IDIC^{BP}, the better predictive performance of the candidate model using the Bayesian predictive distribution.*

Remark 4.8 *According to Li, et al (2017), the Bayesian prediction distribution $p^{BP}\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ has smaller risk than the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}\right)$ asymptotically. From Theorem 4.1 in Li, et al (2017) and Corollary 4.3, we can conclude that when n goes to infinity, the risk for the model chosen by IDIC is equivalent to that by DIC$_1$ and the risk of the model chosen by IDIC^{BP} is equivalent to that by DIC$_1^{BP}$. However, the model chose by IDIC^{BP} yields a smaller risk than that by IDIC asymptotically. Thus, IDIC^{BP} perform better than IDIC in choosing a model to make prediction. Furthermore, DIC$_1^{BP}$ and IDIC^{BP} perform equivalently in choosing a model to make prediction. Clearly, IDIC and IDIC^{BP} are computationally tractable alternatives to DIC$_1$ and DIC$_1^{BP}$ for comparing latent variable models after MCMC output is available.*

4.3 Computing IDIC and IDIC^{BP} for latent variable models

Since IDIC and IDIC^{BP} have nearly identical expressions with a small difference in the penalty terms, knowing one of them implies that the other is automatically known. For this reason, we focus on the computational issue of IDIC in this section. IDIC has two terms, $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ and P_D^I. When $\ln p(\mathbf{y}|\boldsymbol{\theta})$ does not have an analytical expression, both $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ and P_D^I are difficult to compute.

To calculate IDIC, one needs to calculate $p(\mathbf{y}|\boldsymbol{\theta})$ and its derivatives with respect to $\boldsymbol{\theta}$ (but there is no need to optimize $p(\mathbf{y}|\boldsymbol{\theta})$). Since there is no analytical expression for $p(\mathbf{y}|\boldsymbol{\theta})$ for many latent variable models, in this section, we show how to use the EM algorithm, the Kalman filter, and the particle filters to calculate $p(\mathbf{y}|\boldsymbol{\theta})$ and its derivatives with respect to $\boldsymbol{\theta}$.

4.3.1 Computing IDIC by the EM algorithm

In this subsection 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 IDIC for the latent variable models. The EM algorithm is a powerful tool to deal with latent variable models. Instead of maximizing the observed-data likelihood function, the EM algorithm maximizes the so-called $\mathcal{Q}$ function given by

$$\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)}) = E_{\boldsymbol{\theta}^{(r)}}\{\mathcal{L}_c(\mathbf{y},\boldsymbol{z}|\boldsymbol{\theta})|\mathbf{y},\theta^{(r)}\}, \tag{19}$$

where $\mathcal{L}_c(\mathbf{y},\boldsymbol{z}|\boldsymbol{\theta}) := p(\mathbf{y},\boldsymbol{z}|\boldsymbol{\theta})$ is the complete-data likelihood function. The $\mathcal{Q}$function is the conditional expectation of $\mathcal{L}_c(\mathbf{y},\boldsymbol{z}|\boldsymbol{\theta})$ with respect to the conditional distribution $p(\boldsymbol{z}|\mathbf{y},\boldsymbol{\theta}^{(r)})$ where $\boldsymbol{\theta}^{(r)}$ is a current fit of the parameter. The EM algorithm consists of two steps: the *expectation* (E) step and the *maximization* (M) step. The E-step evaluates $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$. The M-step determines a $\boldsymbol{\theta}^{(r)}$ that maximizes $\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^{(r)})$. Under some mild regularity conditions, for large enough r, $\{\boldsymbol{\theta}^{(r)}\}$ obtained from the EM algorithm is the MLE, $\widehat{\boldsymbol{\theta}}$. For more details about the EM algorithm, see Dempster et al. (1977).

Although the EM algorithm is a good approach to dealing with latent variable models, the numerical optimization in the M-step is often unstable. Not surprisingly, the EM algorithm has been less popular to estimate latent variables models compared with the MCMC techniques. However, we will show that, without using the numerical optimization in the M-step, the theoretical properties of the EM algorithm facilitate computation of IDIC for latent variable models.

It is noted that for any $\boldsymbol{\theta}$ and $\boldsymbol{\theta}^{*}$in Θ, let $\mathcal{H}(\boldsymbol{\theta}|\boldsymbol{\theta}^{*}) = \int \ln p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta}^{*})d\mathbf{z}$, the so-called $\mathcal{H}$ function in the EM algorithm. It was shown in that

$$\ln p(\mathbf{y}|\boldsymbol{\theta}) = \mathcal{Q}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{*}\right) - \mathcal{H}\left(\boldsymbol{\theta}|\boldsymbol{\theta}^{*}\right).$$

Hence, $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ may be obtained as

$$\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}}) = \mathcal{Q}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}) - \mathcal{H}(\bar{\boldsymbol{\theta}}|\bar{\boldsymbol{\theta}}). \tag{20}$$

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

For the second term in (20), 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 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 when $i\neq j$, and $\mathbf{z}_i$ is low-dimensional, say ≤ 5, then a nonparametric approach may be used to approximate $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$. Note that

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

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 $p\left(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta}\right)$. Using these random samples, one can then use nonparametric techniques such as the kernel-based methods to approximate $p\left(\mathbf{z}_i|\mathbf{y}_i,\boldsymbol{\theta}\right)$. 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 $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})=-1448.97$, IDIC= 2901.46 for Model 1 and $\ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})=-1449.41$, IDIC= 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 DIC$_7$ = 2884.37 for Model 1 and DIC$_7$ = 2852.85 for Model 2. The highly distinctive difference between them suggests that DIC$_7$ is not a reliable model selection criterion for the model. Note that DIC$_1$ is very difficult 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 of models, 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, $p(\mathbf{z}|\mathbf{y},\boldsymbol{\theta})$, may be expressed as,

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

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^I, we have to calculate the second derivative of the observed-data likelihood function in P_D^I. Under the mild regularity condition, Louis (1982) showed that this second derivative may be expressed as:

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

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 $\boldsymbol{\theta}$.

If $\mathcal{Q}$ function has an analytical expression, Oakes (1999) showed that the second derivative has an equivalent expression

$$\mathbf{I}(\boldsymbol{\theta}) = -\frac{\partial^2\mathcal{L}_o(\mathbf{y}|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = \left\{-\frac{\partial^2\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{\partial^2\mathcal{Q}(\boldsymbol{\theta}|\boldsymbol{\theta}^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{*\prime}}\right\}_{\boldsymbol{\theta}^*=\boldsymbol{\theta}}. \quad (22)$$

If the analytical $\mathcal{Q}$ function not available, we may approximate the second derivatives by,

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

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

Although EM algorithm is a very general approach to analyzing latent variable models, it is very cumbersome to deal with dynamic latent variable models, such as, state space models because we have to compute the derivatives recursively (Doucet and Shephard, 2012). Alternatively, one can compute IDIC using the Kalman filter and particle filters.

4.3.2 Computing IDIC by the Kalman filter

In economics, many time series models can be represented by a linear Gaussian state space form. The Kalman filter is an efficient recursive method for computing the optimal linear forecasts in such models. It also gives the exact likelihood function of the model. Here, we only present the basic idea of the Kalman filter for analyzing liner state space models. One may refer to Harvey (1989) for the detailed textbook treatment.

Consider a general linear state space model,

$$\begin{aligned} z_t &= Tz_{t-1} + R\varepsilon_t, \\ y_t &= D + Cz_t + \xi_t, \end{aligned}$$

where $\varepsilon_t \sim N(0, Q)$, $\xi_t \sim N(0, H)$, T is $n_s \times n_s$, R is $n_s \times n_e$, D is $n \times 1$, C is $n \times n_s$, Q is $n_e \times n_e$, H is $n \times n$. These six coefficient matrices are functions of a vector of parameters $\boldsymbol{\theta}$ which is $n_q \times 1$.

Let $\mathbf{y}^s = (y_1, y_2, \ldots, y_s)$, $z_t^s = E(z_t|\mathbf{y}^s)$, $\Sigma_t^s = E\{(z_t - z_t^s)(z_t - z_t^s)'|\mathbf{y}^s\}$. With the initial conditions, z_0^0 and Σ_0^0, for $t = 1, 2, \ldots, n$, the Kalman filter recursively implements the following steps

$$\begin{aligned} z_t^{t-1} &= Tz_{t-1}^{t-1}, \\ \Sigma_t^{t-1} &= T\Sigma_{t-1}^{t-1}T' + RQR', \end{aligned}$$

and

$$\begin{aligned} z_t^t &= z_t^{t-1} + K_t\left(y_t - D - Cz_t^{t-1}\right), \\ \Sigma_t^t &= \left[I_{n_s} - K_tC\right]\Sigma_t^{t-1}, \end{aligned}$$

where

$$K_t = \Sigma_t^{t-1}C'\left[C\Sigma_t^{t-1}C' + H\right]^{-1}.$$

The observed-data log-likelihood is given by

$$\begin{aligned} \ln p(\mathbf{y}|\boldsymbol{\theta}) &= -\sum_{t=1}^{n}\left[\frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|F_t| + \frac{1}{2}\left(y_t - D - Cz_t^{t-1}\right)' F_t^{-1}\left(y_t - D - Cz_t^{t-1}\right)\right] \\ &= -\sum_{t=1}^{n}\left[\frac{n}{2}\ln 2\pi + \frac{1}{2}\ln|F_t| + \frac{1}{2}\omega_t' F_t^{-1}\omega_t\right], \end{aligned}$$

where $F_t = CP_t^{t-1}C' + H$, $\omega_t = y_t - D - Cz_t^{t-1}$. Clearly, $\ln p(y|\boldsymbol{\theta})$ has to be calculated recursively since F_t and z_t^{t-1} are only available recursively. Similarly, $s_t(\boldsymbol{\theta})$ and $h_t(\boldsymbol{\theta})$ has to be computed recursively. To calculate $s_t(\boldsymbol{\theta})$ and $h_t(\boldsymbol{\theta})$, we need to calculate the first and second order derivatives of $|F_t|$, $\omega_t' F_t^{-1}\omega_t$ recursively. For details, one can refer to Iskrev (2008) and Herbst (2010).

4.3.3 Computing IDIC by particle filters

In practice, the nonlinear non-Gaussian state space models have been widely used in empirical works but they cannot be analyzed using the Kalman filter. Instead, one can use another class of recursive filtering algorithms known as particle filters. We only present the basic idea of particle filters here and refer the reader to recent review papers on particle filters by Doucet and Johansen (2009) and Creal (2012) for greater details.

Let $z_{t+1}|z_t \sim f\left(z_{t+1}|z_t, \boldsymbol{\theta}\right)$ and $y_t|z_t \sim g\left(y_t|z_t, \boldsymbol{\theta}\right)$. Let the initial density of z be $\mu\left(z|\boldsymbol{\theta}\right)$. The joint density of $\left(\mathbf{z}^t, \mathbf{y}^t\right)$ is

$$p\left(\mathbf{z}^t, \mathbf{y}^t|\boldsymbol{\theta}\right) = \mu\left(z_1|\boldsymbol{\theta}\right) \prod_{k=2}^{t} f\left(z_k|z_{k-1}, \boldsymbol{\theta}\right) \prod_{k=1}^{t} g\left(y_k|z_k, \boldsymbol{\theta}\right),$$

and hence

$$p\left(\mathbf{y}^t|\boldsymbol{\theta}\right) = \int p\left(\mathbf{z}^t, \mathbf{y}^t|\boldsymbol{\theta}\right) d\mathbf{z}^t.$$

For nonlinear non-Gaussian state space models, neither $p\left(\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right)$ nor $p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)$ are available in closed-form. The goal here is to calculate $p\left(\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right)$, $p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)$, and $\mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta})$ sequentially for $t = 1, \ldots, n$. The idea of the using particle filters is to approximate $p\left(\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right) d\mathbf{z}^t$ by its empirical measure. An example of particle filters is the Sequential Important Sampling and Resampling (SISR) algorithm which iterates the following step for $i = 1, \ldots, N$,

Step 1: At $t = 1$, $z_1^{(i)} \sim \mu\left(\cdot\right)$,

$$w_1\left(\mathbf{z}^{1(i)}\right) = \frac{\mu\left(z_1^{(i)}|\boldsymbol{\theta}\right) g\left(y_1|z_1^{(i)}, \boldsymbol{\theta}\right)}{q_1\left(z_1^{(i)}\right)}, \quad W_1^{(i)} = \frac{w_1\left(\mathbf{z}^{1(i)}\right)}{\sum_{i=1}^{N} w_1\left(\mathbf{z}^{1(i)}\right)},$$

$\mathbf{z}^{1(i)} = z_1^{(i)}$. Resample $\left(W_1^{(i)}, \mathbf{z}^{1(i)}\right)$ to obtain new particles $\left(\frac{1}{N}, \widetilde{\mathbf{z}}^{1(i)}\right)$.

Step 2: At $t \geq 2$, $z_t^{(i)} \sim q_n\left(\cdot|\widetilde{\mathbf{z}}^{t-1(i)}\right)$,

$$w_t\left(\mathbf{z}^{t(i)}\right) = \frac{f\left(z_t^{(i)}|\widetilde{z}_{t-1}^{(i)}, \boldsymbol{\theta}\right) g\left(y_t|\widetilde{z}_t^{(i)}, \boldsymbol{\theta}\right)}{q_t\left(z_t^{(i)}|\widetilde{\mathbf{z}}^{t-1(i)}\right)}, \quad W_t^{(i)} = \frac{w_t\left(\mathbf{z}^{t(i)}\right)}{\sum_{i=1}^{N} w_t\left(\mathbf{z}^{t(i)}\right)},$$

$\mathbf{z}^{t(i)} = \left(\widetilde{\mathbf{z}}^{t-1(i)}, z_t^{(i)}\right)$. Resample $\left(W_t^{(i)}, \mathbf{z}^{t(i)}\right)$ to obtain new particles $\left(\frac{1}{N}, \widetilde{\mathbf{z}}^{t(i)}\right)$.

Step 3: Approximate the conditional distribution $p_{\boldsymbol{\theta}}\left(d\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right)$ by its empirical measure

$$\widehat{p}\left(d\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right) = \sum_{i=1}^{N} W_t^{(i)} \delta_{\mathbf{z}^{t(i)}}\left(d\mathbf{z}^t\right) \quad \text{or} \quad \widetilde{p}_{\boldsymbol{\theta}}\left(d\mathbf{z}^t|\mathbf{y}^t, \boldsymbol{\theta}\right) = \frac{1}{N} \sum_{i=1}^{N} \delta_{\widetilde{\mathbf{z}}^{t(i)}}\left(d\mathbf{z}^t\right),$$

and

$$\widehat{p}\left(y_t|\mathbf{y}^{t-1}, \boldsymbol{\theta}\right) = \frac{1}{N} \sum_{i=1}^{N} w_t\left(\mathbf{z}^{t(i)}\right),$$

where N is the number of particles and $q_t(\cdot|\cdot)$ is the proposal density.

With the empirical measure $\left\{\widehat{p}\left(d\mathbf{z}^t|\mathbf{y}^t,\boldsymbol{\theta}\right)\right\}_{t=1:n}$, we can approximate the integral

$$I_t = \int \varphi_t\left(\mathbf{z}^t\right) p\left(\mathbf{z}^t|\mathbf{y}^t,\boldsymbol{\theta}\right) d\mathbf{z}^t,$$

by

$$\widehat{I}_t = \int \varphi_t\left(\mathbf{z}^t\right) \widehat{p}\left(d\mathbf{z}^t|\mathbf{y}^t,\boldsymbol{\theta}\right) = \sum_{i=1}^{N} W_t^{(i)} \varphi_t\left(\mathbf{z}^{t(i)}\right),$$

for $t = 1, \cdots, n$, where $\varphi_t\left(\mathbf{z}^t\right)$ is the target function. If one chooses $\varphi_t\left(\mathbf{z}^t\right) = \partial \ln p\left(\mathbf{z}^t, \mathbf{y}^t|\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}$, then it is easy to show that

$$\mathbf{s}(\mathbf{y}^t,\boldsymbol{\theta}) = \int \frac{\partial \ln p\left(\mathbf{z}_t,\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} p\left(\mathbf{z}_t|\mathbf{y}^t,\boldsymbol{\theta}\right) d\mathbf{z}_t,$$

$$-\mathbf{H}(\mathbf{y}^t,\boldsymbol{\theta}) = \mathbf{s}(\mathbf{y}^t,\boldsymbol{\theta})\mathbf{s}(\mathbf{y}^t,\boldsymbol{\theta})' - \frac{\partial^2 p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}{p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)}$$

where

$$\begin{aligned}\frac{\partial^2 p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}{p\left(\mathbf{y}^t|\boldsymbol{\theta}\right)} &= \int \frac{\partial \ln p\left(\mathbf{z}_t,\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}} \frac{\partial \ln p\left(\mathbf{z}_t,\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}' p\left(\mathbf{z}_t|\mathbf{y}^t,\boldsymbol{\theta}\right) d\mathbf{z}_t \\ &\quad + \int \frac{\partial^2 \ln p\left(\mathbf{z}_t,\mathbf{y}^t|\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} p\left(\mathbf{z}_t|\mathbf{y}^t,\boldsymbol{\theta}\right) d\mathbf{z}_t,\end{aligned}$$

by the Fisher and Louis identities that are based only on the marginal density $p\left(\mathbf{z}_t|\mathbf{y}^t,\boldsymbol{\theta}\right)$ (Poyiadjis and Doucet, 2011). Therefore, $s(y^t,\boldsymbol{\theta})$ and $H(y^t,\boldsymbol{\theta})$ can be obtained recursively.

Based on different proposal density $q_t(\cdot|\cdot)$, different particle filtering algorithms have been proposed in the literature, including the bootstrap particle filters of Gordon et al. (1993) and the auxiliary particle filters of Pitt and Shephard (1999). In this paper, we use the auxiliary particle filters to compute $s(y^t,\boldsymbol{\theta})$, $H(y^t,\boldsymbol{\theta})$. The details about how to compute $s(y^t,\boldsymbol{\theta})$ and $H(y^t,\boldsymbol{\theta})$ using the particle filters can be found in Poyiadjis and Doucet (2011) and Doucet and Shephard (2012).

5 Applications

We now illustrate the proposed method in three applications. The first example is asset pricing models under the Student t distribution. The likelihood functions of these models not only have analytical form, but also can be rewritten in a latent variable form. We choose this example to compare the two alternative formulations of the same model, paying particular attention to the impact the two equivalent formulations on DIC and IDIC. In the second example linear state space models are considered. In this case $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is not available in closed-form, but the Kalman filter provides a recursive algorithm to evaluate it. In the third example, $p(\mathbf{y}|\bar{\boldsymbol{\theta}})$ is not available in closed-form and the Kalman filter is not applicable. Given that DIC_1 is too difficult to compute, we calculate IDIC using the proposed method.

5.1 Factor asset pricing models

Factor asset pricing models are important in modern finance. There models generally assume that the return distribution is normal. Unfortunately, there has been overwhelming empirical evidence against normality for asset returns, which have led researchers to investigate asset pricing models with heavy-tailed distributions. Zhou (1993) and Kan and Zhou (2003) suggested using the multivariate t distribution to replace the multivariate normal distribution. Moreover, based on the efficient market theory, the asset excess premium should not be statistically different from zero. At last, the multivariate t distribution can be rewritten as scale-mixture framework to become a latent variable model. Hence, we consider the following six asset pricing models:

$$
\begin{aligned}
&\text{Model 1:} R_t = \boldsymbol{\beta'}\boldsymbol{F}_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \boldsymbol{\Sigma}],\\
&\text{Model 2:} R_t = \alpha + \boldsymbol{\beta'}\boldsymbol{F}_t + \epsilon_t, \epsilon_t \sim N[\mathbf{0}, \boldsymbol{\Sigma}],\\
&\text{Model 3:} R_t = \boldsymbol{\beta'}\boldsymbol{F}_t + \epsilon_t, \epsilon_t \sim t[\mathbf{0}, \boldsymbol{\Sigma}, \nu],\\
&\text{Model 4:} R_t = \boldsymbol{\beta'}\boldsymbol{F}_t + \boldsymbol{\epsilon}_t, \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right),\\
&\text{Model 5:} R_t = \boldsymbol{\alpha} + \boldsymbol{\beta'}\boldsymbol{F}_t + \epsilon_t, \epsilon_t \sim t[\mathbf{0}, \boldsymbol{\Sigma}, \nu],\\
&\text{Model 6:} R_t = \boldsymbol{\alpha} + \boldsymbol{\beta'}\boldsymbol{F}_t + \epsilon_t, \boldsymbol{\epsilon}_t \sim N(\mathbf{0}, \boldsymbol{\Sigma}/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2}, \frac{\nu}{2}\right),
\end{aligned}
$$

where R_t is the excess return of portfolio at period t with $N \times 1$ dimension, $\boldsymbol{F}_t$ a $K \times 1$ vector of factor portfolio excess returns, $\boldsymbol{\alpha}$ a $N \times 1$ vector of intercepts, $\boldsymbol{\beta}$ a $N \times K$ vector of scaled covariances, ϵ_t the random error, $t = 1, 2, \cdots, n$. For convenience, we restrict $\boldsymbol{\Sigma}$ to be a diagonal matrix and ν to be a known constant as $\nu = 3$. It is noted that Model 4 is the scale-mixture distributional representation of Model 3, and Model 5 is the scale mixture distributional representation of Model 6.

Monthly returns of 25 portfolios, constructed at the end of each June, are the intersections of 5 portfolios formed on size (market equity, ME) and 5 portfolios formed on the ratio of book equity to market equity (BE/ME). The Fama/French's three factors, market excess return, SMB (Small Minus Big), HML (High Minus Low) are used as the explanatory factors (Fama and French, 1993). The sample period is from July 1926 to November 2017, so that $N = 25$, $n = 1097$. The data are freely available from the data library of Kenneth French.[3]

Bayesian inference for factor asset pricing models has attracted a considerable amount of attentions in the empirical asset pricing literature. Avramov and Zhou (2010) provided an

[3] http://mba.tuck.dartmouth.edu/pages/faculty/ken.french/data_library.html

Table 1: Model selection results for Fama-French three factor models

Model	M_1	M_2	M_3	M_4	M_5	M_6
P	100	125	100	100	125	125
$P_{D,1}$	100	125	100	100	125	125
DIC$_1$	-132196	-132762	-143510	-143510	-144635	-144635
$P_{D,7}$	NA	NA	NA	1090	NA	1115
DIC$_7$	NA	NA	NA	-145159	NA	-146339
P_D^I	100	125	100	100	126	126
IDIC	-132196	-132762	-143509	-143509	-144634	-144634
IDIC^{BP}	-132227	-132800	-143540	-143540	-144672	-144672

excellent review of the literature on Bayesian portfolio analysis. To obtain MCMC output, we need specify the prior distributions for parameters. Here, to represent the prior ignorance, we assign some vague conjugate prior distributions,

$$\alpha_i \sim N[0, 100], \beta_{ij} \sim N[0, 100], \phi_{ii}^{-1} \sim \Gamma[0.01, 0.01].$$

Here, we draw 100,000 random observations from the posterior distributions in each model where the first 40,000 is used as the burn-in sample, and the next 60,000 iterations is collected with every 3^{rd} observation as effective observations. Hence, these are 20,000 effective observations.

To compare these models, based on 20,000 effective observations, we calculate DIC$_1$, the corresponding $P_{D,1}$, IDIC, the corresponding P_D^I, and IDIC^{BP} for each candidate model, and DIC$_7$ and the corresponding P_D (denoted by $P_{D,7}$) for Model 4 and Model 6 as there are latent variables in these two models. The results are reported in Table 1. Several interesting findings emerge from Table 1. First, DIC$_1$ in Model 3 is very different from DIC$_7$ in Model 4 although these two models are the same. The reason for the difference is that in Model 3 there is no latent variable whereas in Model 4 the scale-mixture representation of the Student t distribution introduces latent variables, $\{\omega_t\}$. Due to the difference, the common practice of DIC for Model 3 is DIC$_1$ and for Model 4 is DIC$_7$. The sharp difference between the two DIC values for the identical model is clearly unsatisfactory. For the same reason, DIC$_1$ in Model 5 is very different from DIC$_7$ in Model 6. Second, the asymptotic results developed in Li, et al (2017) and in Theorem 4.1 above suggest that $P_{D,1}$ and P_D^I should be close to the actual number of the parameters, P, if the prior distribution is dominated by the likelihood function. The results are confirmed by Table 1. Not surprisingly, $P_{D,1}$ is almost identical to P_D^I and DIC$_1$ and IDIC are almost the same for each candidate model. Finally, DIC, IDIC, and IDIC^{BP} all pick Model 6 (and Model 5) as the best model.

5.2 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 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), we consider the following 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, Φ 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, Σ is assume to be diagonal and ε_t and η_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 MCMC technique 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:

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

Table 2: Model selection results for dynamic factor models

Model	M_1	M_2	M_3
Number of Parameters	752	1385	2019
$P_{D,7}$	354	971	1404
DIC$_7$	-23288	-37851	-44568
Number of Parameters	241	363	486
P_D^I	88	203	316
IDIC	-22418	-34842	-40383
IDIC^{BP}	-22444	-34901	-40476

where M_0 is a $K \times K$ identity matrix, L_i the ith $(i > K)$ column of L. The diagonal elements of Q_0 are set to be the residual variances of the corresponding AR(1) model, $\{\hat{\sigma}_i^2\}$. The diagonal elements of Ω_0 are constructed so that the prior variance of the parameter on the jth variable in the ith 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 DIC$_7$ and IDIC to compare models with different number 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.

Based on the 20,000 samples, we compute DIC$_7$ and IDIC for all three models. The Kalman filter algorithm is used to approximate the observed-data likelihood at the posterior mean. Table 2 reports the simple count of the number of parameters (including the latent variables), DIC$_7$, the corresponding $P_{D,7}$, the simple count of the number of parameters (P which excludes the latent variables), IDIC, the corresponding P_D^I and IDIC^{BP}. DIC$_7$, IDIC, IDIC^{BP} 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. Since very informative priors have been used, P_D^I is smaller than the actual number of parameters for each candidate model.

5.3 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:

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

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 M_1.

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

$$\mu \sim N(0, 100), \phi \sim Beta(1, 1), \quad 1/\tau^2 \sim \Gamma(0.001, 0.001).$$

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:

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

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, ρ 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 ρ as:

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

Our goal here is to compare the two models using DIC$_7$ and IDIC. In both cases, $p(\mathbf{y}|\boldsymbol{\theta})$ is not available in closed-form. Since both specifications are nonlinear non-Gaussian state space models, the Kalman filter is not applicable, making DIC$_1$ is time consuming to compute. To compute IDIC and IDIC^{BP}, we use a particle filtering algorithm to evaluate the observed-data likelihood and its second derivative.

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 3. Note that the in M_2, the posterior mean of ρ is very close to zero, relative to its posterior standard error.

Table 3: Posterior mean and standard error of parameters in M_1 and M_2

	M_1		M_2	
Parameter	Mean	SE	Mean	SE
μ	-0.6733	0.3282	-0.6485	0.3377
ϕ	0.9733	0.0127	0.9802	0.0138
ρ	NA	NA	-0.0575	0.1570
τ	0.1698	0.0378	0.1661	0.0391

Table 4 reports DIC$_7$, $P_{D,7}$, IDIC, P_D^I and IDIC^{BP}. The following findings can be obtained from Table 3. First and foremost, IDIC and IDIC^{BP} suggest the same ranking of the competing models, but DIC$_7$ is different. In particular, by dropping the value by 43.3 comparing to IDIC, DIC$_7$ suggests that M_2 is better that M_1. According to DIC$_7$, M_1 and M_2 perform nearly the same judged by $D(\bar{\boldsymbol{\theta}})$. However, M_2 reduces the effective number of parameters by 22.3 over M_1. This reduction of the model complexity is the reason why 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, IDIC suggests that M_1 is slightly better that M_2 although the difference is not worth to mention. In IDIC, P_D^I 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. This empirical example clearly demonstrates that IDIC and IDIC^{BP} is a more reliable model selection criterion that DIC$_7$.

6 Conclusion

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 to define the likelihood function for the purpose of obtaining DIC. This is because, although the likelihood function based on data augmentation greatly simplifies calculation of DIC, it makes the number of parameters increases with the number of observations, invalidating the standard Bayesian large sample theory and the ML asymptotic theory, which are needed to show that DIC is an asymptotically unbiased estimator of the expected KL divergence between the DGP and the predictive distributions. In addition, the use of data

Table 4: Model selection results for M_1 and M_2

Model	M_1	M_2
$P_{D,7}$	53.60	31.33
$D(\bar{\boldsymbol{\theta}})$	1695.40	1693.36
DIC$_7$	1802.52	1756.21
P_D^I	2.32	3.24
$D(\bar{\boldsymbol{\theta}})$	1837.81	1837.78
IDIC	1842.50	1844.30
IDIC^{BP}	1841.77	1843.31

augmentation makes DIC very sensitive to nonlinear transformations of latent variables and distributional representations.

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

We introduce integrated deviance information criterion (IDIC) for comparing latent variable models. IDIC is constructed on observed-data likelihood which integrates the latent variable out of complete-data likelihood. We show that IDIC can be justified by the standard Bayesian asymptotic theory. In particular, we show that IDIC is an asymptotically unbiased estimator of the expected KL divergence when the loss function is based on a plug-in predictive distribution. We then develop a simple and general approach to computing IDIC for latent variable models. Since the latent variables are not treated as parameters in defining IDIC, IDIC is robust to nonlinear transformations of the latent variables. Asymptotic justification, computational tractability and robustness to transformation of latent variables are the three main advantages of IDIC. These advantages are illustrated using some popular models in economics and finance.

In addition, based on the Bayesian predictive distribution, another version of IDIC, denoted as IDIC^{BP}, is also developed. It can be shown that IDIC^{BP} is an asymptotically unbiased estimator of the expected KL divergence when the loss function is based on the Bayesian predictive distribution. Furthermore, IDIC^{BP} has a smaller penalty term than the original IDIC. It is invariant to reparametrization and yields a smaller risk than the IDIC asymptotically. It is trivial to compute if IDIC is available.

It should be pointed out that both DIC$_1$ and IDIC require that the candidate models are good models in the sense that they can well approximate the DGP and that the standard ML theory holds true. It is important to relax this assumption to allow the possibility that the candidate models are misspecified asymptotically. This line of research will be pursued in later work.

Appendix

6.1 Proof of Lemma 4.1

In this subsection, for any function $f(\boldsymbol{\theta})$, let $f^{(j)}(\boldsymbol{\theta})$ be the jth order derivative of $f(\boldsymbol{\theta})$ for $j = 1, 2, 3, 4, 5$. Furthermore, let $\widehat{f}$ be the value of function f evaluated at $\widehat{\boldsymbol{\theta}}$, i.e., $\widehat{f} := f\left(\widehat{\boldsymbol{\theta}}\right)$ and for convenience of exposition, we write $\frac{\partial^d}{\partial\theta_{j_1}\partial\theta_{j_2}\cdots\partial\theta_{j_d}} f(\boldsymbol{\theta})$ as $f_{j_1\cdots j_d}$ and let $\widehat{f}_{j_1\cdots j_d} := f_{j_1\cdots j_d}\left(\widehat{\boldsymbol{\theta}}\right)$. For the definition of high order derivatives, we follow Magnus and Neudecker (1999), except that the first order derivative of a scalar function in our setting is a column vector. Then the Hessian matrix at $\boldsymbol{\theta}$ is denoted by $h_n^{(2)}(\boldsymbol{\theta})$ which is briefly written as $h^{(2)}$ and its (i, j)-component is written as h_{ij} while the components of its inverse is written as σ_{ij}. Let μ^4_{ijkq}, μ^6_{ijkqrs}, $\mu^8_{ijkqrstw}$, $\mu^{10}_{ijkqrstwv\beta}$, $\mu^{12}_{ijkqrstwv\beta\tau\phi}$ be the fourth, sixth, eighth, tenth, and twelfth central moments of a multivariate Normal distribution whose covariance matrix is $\widehat{h}^{(-2)} := \left(h^{(2)}(\boldsymbol{\theta})\right)^{-1}|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}}$. In order to prove Lemma 4.1, we first prove two fundamental lemmas and review another lemma.

Lemma 6.1 *For some real-valued function $g(\boldsymbol{\theta})$, if both $(\{h_n(\boldsymbol{\theta})\}, g(\boldsymbol{\theta})b_D(\boldsymbol{\theta}))$ and $(\{h_n(\boldsymbol{\theta})\}, b_D(\boldsymbol{\theta}))$ satisfy the analytical assumptions for the stochastic Laplace method on $\wp_\theta$, then*

$$\frac{\int g(\boldsymbol{\theta})\, b_D(\boldsymbol{\theta}) \exp(-nh_n(\boldsymbol{\theta}))\, d\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta}) \exp(-nh_n(\boldsymbol{\theta}))\, d\boldsymbol{\theta}} = \widehat{g} + \frac{1}{n}B_1 + \frac{1}{n^2}(B_2 - B_3) + O_p\left(\frac{1}{n^3}\right),$$

where

$$B_1 = \frac{1}{2}\sum_{ij} \widehat{\sigma}_{ij}\widehat{g}_{ij} + \frac{\sum_{ij}\widehat{\sigma}_{ij}\widehat{b}_{D,j}\widehat{g}_i}{\widehat{b}_D} - \frac{1}{6}\sum_{ijkq}\widehat{h}_{ijk}\mu^4_{ijkq}\widehat{g}_q,$$

$$
\begin{aligned}
B_2 \;=\; & -\frac{1}{120}\sum_{ijkqrs}\widehat{h}_{ijkqr}\mu^6_{ijkqrs}\widehat{g}_s+\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrst}\mu^8_{ijkqrstw}\widehat{g}_w \\
& -\frac{1}{1296}\sum_{ijkqrstwv\beta}\widehat{h}_{ijk}\widehat{h}_{qrs}\widehat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\widehat{g}_\beta-\frac{1}{24}\frac{\sum_{ijkqrs}\widehat{h}_{ijkq}\mu^6_{ijkqrs}\widehat{b}_{D,s}\widehat{g}_r}{\widehat{b}_D} \\
& +\frac{1}{72}\frac{\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^8_{ijkqrstw}\widehat{b}_{D,w}\widehat{g}_t}{\widehat{b}_D}-\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\widehat{b}_{D,\eta\xi}\widehat{g}_\zeta}{\widehat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\widehat{b}_{D,\eta\xi\omega}\widehat{g}_\zeta}{\widehat{b}_D}-\frac{1}{48}\sum_{ijkqrs}\widehat{h}_{ijkq}\mu^6_{ijkqrs}\widehat{g}_{rs} \\
& +\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^8_{ijkqrstw}\widehat{g}_{tw}-\frac{1}{36}\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\widehat{g}_{\zeta\eta\xi} \\
& +\frac{1}{24}\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\widehat{g}_{\zeta\eta\xi\omega}-\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu^6_{ijk\zeta\eta\xi}\widehat{g}_{\zeta\eta}\widehat{b}_{D,\xi}}{\widehat{b}_D} \\
& +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\widehat{g}_{\zeta\eta\xi}\widehat{b}_{D,\omega}}{\widehat{b}_D}+\frac{1}{4}\frac{\sum_{\zeta\eta\xi\omega}\mu^4_{\zeta\eta\xi\omega}\widehat{g}_{\zeta\eta}\widehat{b}_{D,\xi\omega}}{\widehat{b}_D},
\end{aligned}
$$

$$B_3 = B_4 \times B_1,$$

$$B_4=\frac{1}{2}\sum_{ij}\widehat{\sigma}_{ij}\frac{\widehat{b}_{D,ij}}{\widehat{b}_D}-\frac{1}{6}\sum_{ijkq}\widehat{h}_{ijk}\mu^4_{ijkq}\frac{\widehat{b}_{D,q}}{\widehat{b}_D}+\frac{1}{72}\sum_{ijkqrs}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^6_{ijkqrs}-\frac{1}{24}\sum_{ijkq}\widehat{h}_{ijkq}\mu^4_{ijkq}.$$

Lemma 6.2 *Suppose A is a $P\times P$ matrix, then*

$$\left[vec\,(A)'\otimes\mathbf{I}_P\right]\left[\mathbf{I}_P\otimes vec\,(I_P)\right]=A. \tag{23}$$

Proof. The matrix A has P^2 elements denoted a_{ij}, $i,j=1,2,\dots,P$. Let $e_1,e_2,\dots,e_P$ denote the columns of $P\times P$ identity matrix $\mathbf{I}_P$. We can express A as $A=\sum_{ij}a_{ij}e_ie_j'$, then

$$
\begin{aligned}
& \left[vec\,(A)'\otimes\mathbf{I}_P\right]\left[\mathbf{I}_P\otimes vec\,(I_P)\right] \\
= & \sum_{ij}a_{ij}\left[\left(vec\left(e_ie_j'\right)\otimes\mathbf{I}_P\right)\left(\mathbf{I}_P\otimes vec\,(\mathbf{I}_P)\right)\right]=\sum_{ij}a_{ij}\left[\left(e_j'\otimes e_i'\otimes\mathbf{I}_P\right)\left(\mathbf{I}_P\otimes vec\,(\mathbf{I}_P)\right)\right] \\
= & \sum_{ij}a_{ij}\left[\left(e_j'\mathbf{I}_P\right)\otimes\left(\left(e_i'\otimes\mathbf{I}_P\right)vec\,(\mathbf{I}_P)\right)\right]=\sum_{ij}a_{ij}\left[\left(e_j'\mathbf{I}_P\right)\otimes vec\,(\mathbf{I}_P\mathbf{I}_Pe_i)\right] \\
= & \sum_{ij}a_{ij}\left[e_j'\otimes e_i\right]=\sum_{ij}a_{ij}e_ie_j'=A.
\end{aligned}
$$

The third equality above follows from

$$(B\otimes C)\,(D\otimes E)=BD\otimes CE \tag{24}$$

for four matrices B, C, D and E if BD and CE exist and the fourth equality is because of

$$vec\,(BCD)=(D\otimes B)\,vec\,(C) \tag{25}$$

for three matrices B, C and D if the product BCD is defined. ■

Lemma 6.3 (The Generalized Isserlis Theorem) *If $A = \{\alpha_1, \dots, \alpha_{2N}\}$ is a set of integers such that $1 \leq \alpha_i \leq P$, for each $i \in [1, 2N]$ and $X \in R^P$ is a zero mean multivariate normal random vector then*

$$EX_A = \underset{A}{\Sigma\Pi} E\left(X_i X_j\right), \tag{26}$$

where the notation $\Sigma\Pi$ means summing over all distinct ways of partitioning $X_{\alpha_1}, \dots, X_{\alpha_{2N}}$ into pairs (X_i, X_j) and each summand is the product of the N pairs. This yields $(2N)!/\left(2^N N!\right) = (2N-1)!!$ terms in the sum where $(2N-1)!!$ is the double factorial such that $(2N-1)!! = (2N-1)(2N-3)\dots 1$.

The Isserlis theorem, first obtained by Isserlis (1918), expresses the higher order moments of a zero mean Gaussian vector in terms of its covariance matrix. The generalized Isserlis theorem is due to Withers (1985) and Vignat (2012). On the basis of Lemma 6.1, 6.2 and 6.3, in the following, we prove the Lemma 4.1.

Proof. First, we define a function $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, and each element of $\mathbf{g}(\boldsymbol{\theta})$ is given as $g_z(\boldsymbol{\theta}) = \boldsymbol{\theta}_z$, $z = 1, \dots, P$. Denote $\mathbf{g}^{(1)}$, a $P \times P$ matrix, is the first order derivative of $\mathbf{g}$ evaluated at $\boldsymbol{\theta}$ and $\mathbf{g}^{(1)}_{\cdot z}$ is the z_{th} column of $\mathbf{g}^{(1)}$. It is noted that since $\mathbf{g}(\boldsymbol{\theta}) = \boldsymbol{\theta}$, $\mathbf{g}^{(1)} = \mathbf{I}_P$ which is $P \times P$ identity matrix.

For $z = 1, \dots, P$, $g_z(\boldsymbol{\theta})$ is a real-valued function. Hence, using Lemma 6.1, we can get that for each z

$$\frac{\int g_z(\boldsymbol{\theta}) b_D(\boldsymbol{\theta}) \exp\left(-n h_n(\boldsymbol{\theta})\right) d\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta}) \exp\left(-n h_n(\boldsymbol{\theta})\right) d\boldsymbol{\theta}} = g_z(\boldsymbol{\theta}_n) + \frac{1}{n} B^1_{1,z} + \frac{1}{n^2}\left(B^1_{2,z} - B^1_{3,z}\right) + O_p\left(\frac{1}{n^3}\right),$$

Then, in the matrix form, we get

$$\frac{\int \mathbf{g}(\boldsymbol{\theta}) b_D(\boldsymbol{\theta}) \exp\left(-n h_n(\boldsymbol{\theta})\right) d\boldsymbol{\theta}}{\int b_D(\boldsymbol{\theta}) \exp\left(-n h_n(\boldsymbol{\theta})\right) d\boldsymbol{\theta}} = \mathbf{g}(\widehat{\boldsymbol{\theta}}) + \frac{1}{n} B^1_1 + \frac{1}{n^2}\left(B^1_2 - B^1_3\right) + O_p\left(\frac{1}{n^3}\right).$$

For each z, note that $g_{z,ij} = \frac{\partial g^2_z(\boldsymbol{\theta})}{\partial\theta\partial\theta'}|_{ij} = \mathbf{0}_{ij}$. Following Lemma 6.1, we have

$$B^1_{1,z} = 0 + \sum_{ij} \widehat{g}_{z,i} \widehat{\sigma}_{ij} \frac{\widehat{b}_{D,j}}{\widehat{b}_D} - \frac{1}{6} \sum_{ijkq} \widehat{h}_{ijk} \mu^4_{ijkq} \widehat{g}_{z,q}.$$

Thus, in the matrix form, we have

$$\begin{aligned} B^1_1 &= \sum_{ij} \widehat{\mathbf{g}}^{(1)}_{\cdot i} \widehat{\sigma}_{ij} \frac{\widehat{b}_{D,j}}{\widehat{b}_D} - \frac{1}{2} \sum_{ijkq} \widehat{\mathbf{g}}^{(1)}_{\cdot q} \widehat{h}_{ijk} \widehat{\sigma}_{ij} \widehat{\sigma}_{kq} = \sum_{ij} \widehat{\mathbf{g}}^{(1)}_{\cdot i} \widehat{\sigma}_{ij} \frac{\widehat{b}_{D,j}}{\widehat{b}_D} - \frac{1}{2} \sum_{ijkq} \widehat{\mathbf{g}}^{(1)}_{\cdot q} \widehat{\sigma}_{qk} \widehat{h}_{ijk} \widehat{\sigma}_{ij} \\ &= \widehat{\mathbf{g}}^{(1)} \widehat{h}^{(-2)} \frac{\widehat{b}^{(1)}_D}{\widehat{b}_D} - \frac{1}{2} \widehat{\mathbf{g}}^{(1)} \widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right), \end{aligned} \tag{27}$$

Hence, we get

$$B_1^1 = \widehat{h}^{(-2)} \frac{\widehat{b}_D^{(1)}}{\widehat{b}_D} - \frac{1}{2} \widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right). \tag{28}$$

Furthermore, for each z

$$\begin{aligned} B_{2,z}^1 \quad = \quad & -\frac{1}{120} \sum_{ijkqrs} \widehat{h}_{ijkqr} \mu_{ijkqrs}^6 \widehat{g}_{z,s} + \frac{1}{144} \sum_{ijkqrstw} \widehat{h}_{ijk} \widehat{h}_{qrst} \mu_{ijkqrstw}^8 \widehat{g}_{z,w} \\ & -\frac{1}{1296} \sum_{ijkqrstwv\beta} \widehat{h}_{ijk} \widehat{h}_{qrs} \widehat{h}_{twv} \mu_{ijkqrstwv\beta}^{10} \widehat{g}_{z,\beta} - \frac{1}{24} \frac{\sum_{ijkqrs} \widehat{h}_{ijkq} \mu_{ijkqrs}^6 \widehat{b}_{D,s} \widehat{g}_{z,r}}{\widehat{b}_D} \\ & +\frac{1}{72} \frac{\sum_{ijkqrstw} \widehat{h}_{ijk} \widehat{h}_{qrs} \mu_{ijkqrstw}^8 \widehat{b}_{D,w} \widehat{g}_{z,t}}{\widehat{b}_D} - \frac{1}{12} \frac{\sum_{ijk\zeta\eta\xi} \widehat{h}_{ijk} \mu_{ijk\zeta\eta\xi}^6 \widehat{b}_{D,\eta\xi} \widehat{g}_{z,\zeta}}{\widehat{b}_D} \\ & +\frac{1}{6} \frac{\sum_{\zeta\eta\xi\omega} \mu_{\zeta\eta\xi\omega}^4 \widehat{b}_{D,\eta\xi\omega} \widehat{g}_{z,\zeta}}{\widehat{b}_D}. \end{aligned}$$

Thus, in matrix form we have

$$\begin{aligned} B_2^1 \quad = \quad & -\frac{1}{120} \sum_{ijkqrs} \widehat{g}_{\cdot s} \widehat{h}_{ijkqr} \mu_{ijkqrs}^6 + \frac{1}{144} \sum_{ijkqrstw} \widehat{g}_{\cdot w} \widehat{h}_{ijk} \widehat{h}_{qrst} \mu_{ijkqrstw}^8 \\ & -\frac{1}{1296} \sum_{ijkqrstwv\beta} \widehat{g}_{\cdot\beta} \widehat{h}_{ijk} \widehat{h}_{qrs} \widehat{h}_{twv} \mu_{ijkqrstwv\beta}^{10} - \frac{1}{24} \frac{\sum_{ijkqrs} \widehat{g}_{\cdot r} \widehat{h}_{ijkq} \mu_{ijkqrs}^6 \widehat{b}_{D,s}}{\widehat{b}_D} \\ & +\frac{1}{72} \frac{\sum_{ijkqrstw} \widehat{g}_{\cdot t} \widehat{h}_{ijk} \widehat{h}_{qrs} \mu_{ijkqrstw}^8 \widehat{b}_{D,w}}{\widehat{b}_D} - \frac{1}{12} \frac{\sum_{ijk\zeta\eta\xi} \widehat{g}_{\cdot\zeta} \widehat{h}_{ijk} \mu_{ijk\zeta\eta\xi}^6 \widehat{b}_{D,\eta\xi}}{\widehat{b}_D} \\ & +\frac{1}{6} \frac{\sum_{\zeta\eta\xi\omega} \widehat{g}_{\cdot\zeta} \mu_{\zeta\eta\xi\omega}^4 \widehat{b}_{D,\eta\xi\omega}}{\widehat{b}_D}. \end{aligned} \tag{29}$$

We can write each item on the right hand side of (29) into matrix form with (26)

$$-\frac{1}{120} \sum_{ijkqrs} \widehat{g}_{\cdot s} \widehat{h}_{ijkqr} \mu_{ijkqrs}^6 = -\frac{1}{8} \sum_{ijkqrs} \widehat{g}_{\cdot s} \widehat{\sigma}_{sr} \widehat{h}_{ijkqr} \widehat{\sigma}_{ij} \widehat{\sigma}_{kq} = -\frac{1}{8} \widehat{g}^{(1)} \widehat{h}^{(-2)} \widehat{h}^{(5)\prime} vec\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right],$$

$$\begin{aligned} & \frac{1}{144} \sum_{ijkqrstw} \widehat{g}_{\cdot w} \widehat{h}_{ijk} \widehat{h}_{qrst} \mu_{ijkqrstw}^8 \\ = \quad & \frac{105}{144} \sum_{ijkqrstw} \widehat{g}_{\cdot w} \widehat{h}_{ijk} \widehat{\sigma}_{ij} \widehat{\sigma}_{kq} \widehat{\sigma}_{rs} \widehat{h}_{qrst} \widehat{\sigma}_{tw} = \frac{35}{48} \sum_{ijkqrstw} \widehat{g}_{\cdot w} \left(\widehat{\sigma}_{wt} \widehat{\sigma}_{rs} \widehat{h}_{\mathbf{tr}sq}\right) \widehat{\sigma}_{qk} \left(\widehat{h}_{kij} \widehat{\sigma}_{ij}\right) \\ = \quad & \frac{35}{48} \widehat{g}^{(1)} \left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right]^{\prime} \widehat{h}^{(4)} \widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right), \end{aligned}$$

$$\begin{aligned}
&-\frac{1}{1296}\sum_{ijkqrstwv\beta}\widehat{g}_{\cdot\beta}\widehat{h}_{ijk}\widehat{h}_{qrs}\widehat{h}_{twv}\mu^{10}_{ijkqrstwv\beta}\\
=\quad&-\frac{945}{1296}\sum_{ijkqrstwv\beta}\widehat{g}_{\cdot\beta}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{kq}\widehat{h}_{qrs}\widehat{\sigma}_{rs}\widehat{\sigma}_{tw}\widehat{h}_{twv}\widehat{\sigma}_{v\beta}\\
=\quad&-\frac{35}{48}\sum_{twv\beta}\widehat{g}_{\cdot\beta}\widehat{\sigma}_{\beta v}\left(\widehat{\sigma}_{tw}\widehat{h}_{twv}\right)\sum_{ijkqrs}\left(\widehat{\sigma}_{ij}\widehat{h}_{ijk}\right)\widehat{\sigma}_{kq}\left(\widehat{h}_{qrs}\widehat{\sigma}_{rs}\right)\\
=\quad&-\frac{35}{48}\widehat{g}^{(1)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\right],
\end{aligned}$$

$$\begin{aligned}
&-\frac{1}{24}\frac{\sum_{ijkqrs}\widehat{g}_{\cdot r}\widehat{h}_{ijkq}\mu^{6}_{ijkqrs}\widehat{b}_{D,s}}{\widehat{b}_D}\\
=\quad&-\frac{15}{24}\sum_{ijkqrs}\widehat{g}_{\cdot r}\widehat{h}_{ijkq}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\widehat{\sigma}_{rs}\frac{\widehat{b}_{D,s}}{\widehat{b}_D}=-\frac{5}{8}\sum_{rs}\widehat{g}_{\cdot r}\widehat{\sigma}_{rs}\frac{\widehat{b}_{D,s}}{\widehat{b}_D}\sum_{ijkq}\widehat{h}_{ijkq}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\\
=\quad&-\frac{5}{8}\widehat{g}^{(1)}\widehat{h}^{(-2)}\frac{\widehat{b}^{(1)}_D}{\widehat{b}_D}\mathbf{tr}\left[\left[\widehat{h}^{(-2)}\otimes vec\left(\widehat{h}^{(-2)}\right)\right]\widehat{h}^{(4)\prime}\right],
\end{aligned}$$

$$\begin{aligned}
&\frac{1}{72}\frac{\sum_{ijkqrstw}\widehat{g}_{\cdot t}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu^{8}_{ijkqrstw}\widehat{b}_{D,w}}{\widehat{b}_D}\\
=\quad&\frac{105}{72}\sum_{ijkqrstw}\widehat{g}_{\cdot t}\widehat{h}_{ijk}\widehat{h}_{qrs}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\widehat{\sigma}_{rs}\widehat{\sigma}_{tw}\frac{\widehat{b}_{D,w}}{\widehat{b}_D}=\frac{35}{24}\sum_{tw}\left(\widehat{g}_{\cdot t}\widehat{\sigma}_{tw}\frac{\widehat{b}_{D,w}}{\widehat{b}_D}\right)\sum_{ijkqrs}\widehat{h}_{ijk}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\widehat{\sigma}_{rs}\widehat{h}_{qrs}\\
=\quad&\frac{35}{24}\widehat{g}^{(1)}\widehat{h}^{(-2)}\frac{\widehat{b}^{(1)}_D}{\widehat{b}_D}\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\right],
\end{aligned}$$

$$\begin{aligned}
&-\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{g}_{\cdot\zeta}\widehat{h}_{ijk}\mu^{6}_{ijk\zeta\eta\xi}\widehat{b}_{D,\eta\xi}}{\widehat{b}_D}\\
=\quad&-\frac{15}{12}\sum_{ijk\zeta\eta\xi}\widehat{g}_{\cdot\zeta}\widehat{h}_{ijk}\widehat{\sigma}_{ij}\widehat{\sigma}_{k\zeta}\widehat{\sigma}_{\eta\xi}\frac{\widehat{b}_{D,\eta\xi}}{\widehat{b}_D}=-\frac{5}{4}\sum_{ijk\zeta}\widehat{g}_{\cdot\zeta}\widehat{\sigma}_{k\zeta}\left(\widehat{h}_{ijk}\widehat{\sigma}_{ij}\right)\sum_{\eta\xi}\widehat{\sigma}_{\eta\xi}\frac{\widehat{b}_{D,\eta\xi}}{\widehat{b}_D}\\
=\quad&-\frac{5}{4}\widehat{g}^{(1)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}^{(2)}_D}{\widehat{b}_D}\right],
\end{aligned}$$

$$\begin{aligned}
\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\widehat{g}_{\cdot\zeta}\mu^{4}_{\zeta\eta\xi\omega}\widehat{b}_{D,\eta\xi\omega}}{\widehat{b}_D}\quad&=\quad\frac{3}{6}\sum_{\zeta\eta\xi\omega}\widehat{g}_{\zeta}\widehat{\sigma}_{\zeta\eta}\widehat{\sigma}_{\xi\omega}\frac{\widehat{b}_{D,\eta\xi\omega}}{\widehat{b}_D}=\frac{1}{2}\sum_{\zeta\eta\xi\omega}\widehat{g}_{\zeta}\widehat{\sigma}_{\zeta\eta}\frac{\widehat{b}_{D,\eta\xi\omega}}{\widehat{b}_D}\widehat{\sigma}_{\xi\omega}\\
&=\quad\frac{1}{2}\widehat{g}^{(1)}\widehat{h}^{(-2)}\frac{\widehat{b}^{(3)\prime}_D}{\widehat{b}_D}\left[vec\left(\widehat{h}^{(-2)}\right)\right].
\end{aligned}$$

Hence, we have

$$\begin{aligned} B_2^1 \;=\; & -\frac{1}{8}\widehat{h}^{(-2)}\widehat{h}^{(5)\prime} vec\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right] + \frac{35}{48}\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right]^{\prime} \widehat{h}^{(4)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right) \\ & -\frac{35}{48}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\right] \\ & -\frac{5}{8}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}\mathbf{tr}\left[\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right]\widehat{h}^{(4)\prime}\right] \\ & +\frac{35}{24}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\right] - \frac{5}{4}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(2)}}{\widehat{b}_D}\right] \\ & +\frac{1}{2}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(3)\prime}}{\widehat{b}_D}\left[vec\left(\widehat{h}^{(-2)}\right)\right]. \end{aligned} \tag{30}$$

For B_3^1, following Lemma 6.1, note that, for any element z, $B_{4,z}^1 = B_4^1$ which is a constant and independent of the element z. We have

$$B_3^1 = B_1^1 \times B_4^1, \tag{31}$$

where

$$B_4^1 = \frac{1}{2}\sum_{ij}\widehat{\sigma}_{ij}\frac{\widehat{b}_{D,ij}}{\widehat{b}_D} - \frac{1}{6}\sum_{ijkq}\widehat{h}_{ijk}\mu_{ijkq}^4\frac{\widehat{b}_{D,q}}{\widehat{b}_D} + \frac{1}{72}\sum_{ijkqrs}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu_{ijkqrs}^6 - \frac{1}{24}\sum_{ijkq}\widehat{h}_{ijkq}\mu_{ijkq}^4. \tag{32}$$

We can write each item on the right hand side of (32) as

$$\frac{1}{2}\sum_{ij}\widehat{\sigma}_{ij}\frac{\widehat{b}_{D,ij}}{\widehat{b}_D} = \frac{1}{2}\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(2)}}{\widehat{b}_D}\right], \tag{33}$$

$$-\frac{1}{6}\sum_{ijkq}\widehat{h}_{ijk}\mu_{ijkq}^4\frac{\widehat{b}_{D,q}}{\widehat{b}_D} = -\frac{3}{6}\sum_{ijkq}\widehat{h}_{ijk}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\frac{\widehat{b}_{D,q}}{\widehat{b}_D} = -\frac{1}{2}vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(-3)}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}, \tag{34}$$

$$\frac{1}{72}\sum_{ijkqrs}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu_{ijkqrs}^6 = \frac{15}{72}\sum_{ijkqrs}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{kq}\widehat{h}_{qrs}\widehat{\sigma}_{rs} = \frac{5}{24}\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\right], \tag{35}$$

$$-\frac{1}{24}\sum_{ijkq}\widehat{h}_{ijkq}\mu_{ijkq}^4 = -\frac{3}{24}\sum_{ijkq}\widehat{h}_{ijkq}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq} = \frac{1}{8}\mathbf{tr}\left[\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right]\widehat{h}^{(4)\prime}\right]. \tag{36}$$

From (32), (33), (34), (35), (36), in the matrix form, we have

$$\begin{aligned} B_4^1 \;=\; & \frac{1}{2}\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(2)}}{\widehat{b}_D}\right] - \frac{1}{2}vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(-3)}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D} + \frac{5}{24}\left[vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\right] \\ & +\frac{1}{8}\mathbf{tr}\left[\left[\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)\right]\widehat{h}^{(4)\prime}\right]. \end{aligned} \tag{37}$$

From (27), (30) and (31), we have

$$\begin{aligned}\overline{\boldsymbol{\theta}} &= \widehat{\boldsymbol{\theta}} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right)\\ &= \widehat{\boldsymbol{\theta}} + \frac{1}{n}B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_4^1 B_1^1\right) + O_p\left(\frac{1}{n^3}\right).\end{aligned}$$

This is the end of proof for the first part of the lemma.

In the following, we prove the second part of the lemma. Define a function $\mathbf{f}\left(\boldsymbol{\theta}\right) = vec\left(\boldsymbol{\theta}\boldsymbol{\theta}'\right)$ which is a $P^2 \times 1$ vector. Hence, we can get the first and second derivatives of $\mathbf{f}$ with respect to $\boldsymbol{\theta}$ as $\mathbf{f}^{(1)}\left(\boldsymbol{\theta}\right) = \boldsymbol{\theta} \otimes \mathbf{I}_P + \mathbf{I}_P \otimes \boldsymbol{\theta}$ and $\mathbf{f}^{(2)}\left(\boldsymbol{\theta}\right) = \left[\left(\mathbf{I}_{P^2} + \mathbf{K}_{PP}\right) \otimes \mathbf{I}_P\right]\left[\mathbf{I}_P \otimes vec\left(\mathbf{I}_P\right)\right]$, where $\mathbf{K}_{mn}$ is a commutation matrix, which is defined by the equation $\mathbf{K}_{mn} vecA = vecA'$ for a $m \times n$ matrix A. If $m = n$, $\mathbf{K}_{mn}$ is simplified as $\mathbf{K}_m$. By the properties of commutation matrix, we have

$$\mathbf{K}_{mn}\left(Y \otimes x\right) = x \otimes Y, \tag{38}$$

$$\left(Y \otimes x'\right)\mathbf{K}_{sm} = x' \otimes Y, \tag{39}$$

where Y is a $n \times s$ matrix, x is a $m \times 1$ vector. Furthermore, for any matrix A_1 and A_2, if A_1 is a $n \times s$ dimensional matrix and A_2 is a $m \times t$ dimensional matrix, then,

$$\mathbf{K}_{mn}\left(A_1 \otimes A_2\right) = \left(A_2 \otimes A_1\right)\mathbf{K}_{ts}. \tag{40}$$

More details about matrix properties, one can refer to Magnus and Neudecker (1979).

Following Lemma 6.1, for each element $f_z(\boldsymbol{\theta})$ which is also real-valued function, we can get that

$$\frac{\int f_z\left(\boldsymbol{\theta}\right) b_D\left(\boldsymbol{\theta}\right)\exp\left(-nh_n\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}}{\int b_D\left(\boldsymbol{\theta}\right)\exp\left(-nh_n\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}} = f_z(\hat{\boldsymbol{\theta}}) + \frac{1}{n}B_{1,z}^2 + \frac{1}{n^2}\left(B_{2,z}^2 - B_{3,z}^2\right) + O_p\left(\frac{1}{n^3}\right).$$

Again, we can rewrite it in the matrix form,

$$\frac{\int \mathbf{f}\left(\boldsymbol{\theta}\right) b_D\left(\boldsymbol{\theta}\right)\exp\left(-nh_n\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}}{\int b_D\left(\boldsymbol{\theta}\right)\exp\left(-nh_n\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}} = \mathbf{f}(\boldsymbol{\theta}) + \frac{1}{n}B_1^2 + \frac{1}{n^2}\left(B_2^2 - B_3^2\right) + O_p\left(\frac{1}{n^3}\right).$$

For each z, we have

$$B_{1,z}^2 = \frac{1}{2}\sum_{ij}\widehat{\sigma}_{ij}\widehat{f}_{z,ij} + \sum_{ij}\widehat{f}_{z,i}\widehat{\sigma}_{ij}\frac{\widehat{b}_{D,j}}{\widehat{b}_D} - \frac{1}{6}\sum_{ijkq}\widehat{h}_{ijk}\mu_{ijkq}^4\widehat{f}_{z,q}.$$

Thus, in the matrix form

$$B_1^2 = \frac{1}{2}\left[\mathbf{I}_{P^2} \otimes vec\left(\widehat{h}^{(-2)}\right)'\right] vec\left(\mathbf{K}_{PP}\widehat{\mathbf{f}}^{(2)}\right) + \sum_{ij}\widehat{\mathbf{f}}_{\cdot i}^{(1)}\widehat{\sigma}_{ij}\frac{\widehat{b}_{D,j}}{\widehat{b}_D} - \frac{1}{2}\sum_{ijkq}\widehat{\mathbf{f}}_{\cdot q}^{(1)}\widehat{h}_{ijk}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}. \tag{41}$$

Note that

$$\begin{aligned}
vec\left(\mathbf{K}_{PP}\widehat{\mathbf{f}}^{(2)}\right) &= vec\left(K_{P^2P}\left[(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\otimes\mathbf{I}_P\right]\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)\right]\right) \\
&= vec\left(\left[\mathbf{I}_P\otimes(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\right]\mathbf{K}_{PP^2}\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)\right]\right) \\
&= \left(\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)\right]'\otimes\left[\mathbf{I}_P\otimes(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\right]\right)vec\left(\mathbf{K}_{PP^2}\right) \\
&= \left(\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)'\otimes\mathbf{I}_P\otimes(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\right)vec\left(\mathbf{K}_{PP^2}\right), \qquad (42)
\end{aligned}$$

where the second equality is due to (40). From (42) and (23), we have

$$\begin{aligned}
&\frac{1}{2}\left[\mathbf{I}_{P^2}\otimes vec\left(\widehat{h}^{(-2)}\right)'\right]vec\left(\mathbf{K}_{PP}\widehat{\mathbf{f}}^{(2)}\right) \\
=\ &\frac{1}{2}\left[\mathbf{I}_{P^2}\otimes vec\left(\widehat{h}^{(-2)}\right)'\right]\left(\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)'\otimes\mathbf{I}_P\otimes(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\right)vec\left(\mathbf{K}_{PP^2}\right) \\
=\ &\frac{1}{2}\left[\left[\mathbf{I}_{P^2}\left(\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)'\otimes\mathbf{I}_P\right)\right]\otimes\left[vec\left(\widehat{h}^{(-2)}\right)'(\mathbf{I}_{P^2}+\mathbf{K}_{PP})\right]\right]vec\left(\mathbf{K}_{PP^2}\right) \\
=\ &\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)'\otimes\mathbf{I}_P\otimes vec\left(\widehat{h}^{(-2)}\right)'\right]vec\left(\mathbf{K}_{PP^2}\right) \\
=\ &vec\left(\left[\mathbf{I}_P\otimes vec\left(\widehat{h}^{(-2)}\right)'\right]\mathbf{K}_{PP^2}\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)\right]\right) \\
=\ &vec\left(\left[vec\left(\widehat{h}^{(-2)}\right)'\otimes\mathbf{I}_P\right]\left[\mathbf{I}_P\otimes vec\left(\mathbf{I}_P\right)\right]\right)=vec\left(\widehat{h}^{(-2)}\right), \qquad (43)
\end{aligned}$$

where the third equality is due to the fact that $vec\left(\widehat{h}^{(-2)}\right)'\mathbf{K}_{PP}=vec\left(\widehat{h}^{(-2)\prime}\right)'=vec\left(\widehat{h}^{(-2)}\right)'$ and the fifth is due to (38). Hence, from (41), (42) and (43),

$$\begin{aligned}
B_1^2 &= vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}-\frac{1}{2}\widehat{\mathbf{f}}^{(1)}\widehat{h}^{(-2)}\widehat{h}^{(-3)\prime}vec\left(\widehat{h}^{(-2)}\right) \\
&= vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}B_1^1. \qquad (44)
\end{aligned}$$

And for each z

$$\begin{aligned}
B_{2,z}^2 &= -\frac{1}{120}\sum_{ijkqrs}\widehat{h}_{ijkqr}\mu_{ijkqrs}^6\widehat{f}_{z,s}+\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrst}\mu_{ijkqrstw}^8\widehat{f}_{z,w} \\
&\quad -\frac{1}{1296}\sum_{ijkqrstwv\beta}\widehat{h}_{ijk}\widehat{h}_{qrs}\widehat{h}_{twv}\mu_{ijkqrstwv\beta}^{10}\widehat{f}_{z,\beta}-\frac{1}{24}\frac{\sum_{ijkqrs}\widehat{h}_{ijkq}\mu_{ijkqrs}^6\widehat{b}_{D,s}\widehat{f}_{z,r}}{\widehat{b}_D} \\
&\quad +\frac{1}{72}\frac{\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu_{ijkqrstw}^8\widehat{b}_{D,w}\widehat{f}_{z,t}}{\widehat{b}_D}-\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu_{ijk\zeta\eta\xi}^6\widehat{b}_{D,\eta\xi}\widehat{f}_{z,\zeta}}{\widehat{b}_D} \\
&\quad +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu_{\zeta\eta\xi\omega}^4\widehat{b}_{D,\eta\xi\omega}\widehat{f}_{z,\zeta}}{\widehat{b}_D}-\frac{15}{48}\sum_{ijkq}\widehat{h}_{ijkq}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\sum_{rs}\widehat{\sigma}_{rs}\widehat{f}_{z,rs} \\
&\quad +\frac{105}{144}\sum_{ijkqrs}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{kq}\widehat{h}_{qrs}\widehat{\sigma}_{rs}\sum_{tw}\widehat{\sigma}_{tw}\widehat{f}_{z,tw}-\frac{15}{12}\sum_{\zeta\eta}\left(\sum_{ijk}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{k\zeta}\sum_{\xi}\widehat{\sigma}_{\eta\xi}\frac{\widehat{b}_{D,\xi}}{\widehat{b}_D}\right)\widehat{f}_{z,\zeta\eta} \\
&\quad +\frac{3}{4}\sum_{\zeta\eta}\widehat{\sigma}_{\zeta\eta}\widehat{f}_{z,\zeta\eta}\sum_{\xi\omega}\widehat{\sigma}_{\xi\omega}\frac{\widehat{b}_{D,\xi\omega}}{\widehat{b}_D}.
\end{aligned}$$

Let $B_{2,z}^2 = B_{21,z}^2 + B_{22,z}^2$ where

$$\begin{aligned}
B_{21,z}^2 &= -\frac{1}{120}\sum_{ijkqrs}\widehat{h}_{ijkqr}\mu_{ijkqrs}^6\widehat{f}_{z,s}+\frac{1}{144}\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrst}\mu_{ijkqrstw}^8\widehat{f}_{z,w} \\
&\quad -\frac{1}{1296}\sum_{ijkqrstwv\beta}\widehat{h}_{ijk}\widehat{h}_{qrs}\widehat{h}_{twv}\mu_{ijkqrstwv\beta}^{10}\widehat{f}_{z,\beta}-\frac{1}{24}\frac{\sum_{ijkqrs}\widehat{h}_{ijkq}\mu_{ijkqrs}^6\widehat{b}_{D,s}\widehat{f}_{z,r}}{\widehat{b}_D} \\
&\quad +\frac{1}{72}\frac{\sum_{ijkqrstw}\widehat{h}_{ijk}\widehat{h}_{qrs}\mu_{ijkqrstw}^8\widehat{b}_{D,w}\widehat{f}_{z,t}}{\widehat{b}_D}-\frac{1}{12}\frac{\sum_{ijk\zeta\eta\xi}\widehat{h}_{ijk}\mu_{ijk\zeta\eta\xi}^6\widehat{b}_{D,\eta\xi}\widehat{f}_{z,\zeta}}{\widehat{b}_D} \\
&\quad +\frac{1}{6}\frac{\sum_{\zeta\eta\xi\omega}\mu_{\zeta\eta\xi\omega}^4\widehat{b}_{D,\eta\xi\omega}\widehat{f}_{z,\zeta}}{\widehat{b}_D},
\end{aligned}$$

and

$$\begin{aligned}
B_{22,z}^2 &= -\frac{15}{48}\sum_{ijkq}\widehat{h}_{ijkq}\widehat{\sigma}_{ij}\widehat{\sigma}_{kq}\sum_{rs}\widehat{\sigma}_{rs}\widehat{f}_{z,rs} \\
&\quad +\frac{105}{144}\sum_{ijkqrs}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{kq}\widehat{h}_{qrs}\widehat{\sigma}_{rs}\sum_{tw}\widehat{\sigma}_{tw}\widehat{f}_{z,tw}-\frac{15}{12}\sum_{\zeta\eta}\left(\sum_{ijk}\widehat{\sigma}_{ij}\widehat{h}_{ijk}\widehat{\sigma}_{k\zeta}\sum_{\xi}\widehat{\sigma}_{\eta\xi}\frac{\widehat{b}_{D,\xi}}{\widehat{b}_D}\right)\widehat{f}_{z,\zeta\eta} \\
&\quad +\frac{3}{4}\sum_{\zeta\eta}\widehat{\sigma}_{\zeta\eta}\widehat{f}_{z,\zeta\eta}\sum_{\xi\omega}\widehat{\sigma}_{\xi\omega}\frac{\widehat{b}_{D,\xi\omega}}{\widehat{b}_D}.
\end{aligned}$$

Then, we rewrite them in the matrix form so that we have

$$B_2^2 = B_{21}^2 + B_{22}^2, \tag{45}$$

where

$$B_{21}^{2} = \widehat{\mathbf{f}}^{(1)} B_2^1 = \left(\widehat{\boldsymbol{\theta}} \otimes \mathbf{I}_P + \mathbf{I}_P \otimes \widehat{\boldsymbol{\theta}}\right) B_2^1 = vec\left(B_2^1 \widehat{\boldsymbol{\theta}}' + \widehat{\boldsymbol{\theta}} B_2^{1\prime}\right), \tag{46}$$

$$\begin{aligned} B_{22}^{2} \;\; = \;\; & -\frac{5}{16} vec\left(\widehat{h}^{(-2)}\right) \mathbf{tr}\left[\left(\widehat{h}^{(-2)} \otimes vec\left(\widehat{h}^{(-2)}\right)'\right) \widehat{h}^{(4)}\right] \\ & + \frac{35}{48} vec\left(\widehat{h}^{(-2)}\right) \left[vec\left(\widehat{h}^{(-2)}\right)' \widehat{h}^{(3)} \widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right)\right] \\ & - \frac{5}{2} vec\left[\widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right) \frac{\widehat{b}_D^{(1)\,\prime}}{\widehat{b}_D} \widehat{h}^{(-2)}\right] \\ & + \frac{3}{4} vec\left(\widehat{h}^{(-2)}\right) \mathbf{tr}\left[\widehat{h}^{(-2)} \frac{\widehat{b}_D^{(2)}}{\widehat{b}_D}\right]. \end{aligned} \tag{47}$$

Since for $z = 1, 2, \cdots, P^2$, by (43), $\sum_{tw} \widehat{\sigma}_{tw} \widehat{f}_{z,tw}$ can be rewritten in the vector form as $vec(\hat{h}^{(-2)})$ and $\frac{1}{2}\sum_{\zeta\eta}\left(\sum_{ijk} \widehat{\sigma}_{ij} \widehat{h}_{ijk} \widehat{\sigma}_{k\zeta} \sum_{\xi} \widehat{\sigma}_{\eta\xi} \frac{\widehat{b}_{D,\xi}}{\widehat{b}_D}\right) \widehat{f}_{z,\zeta\eta}$ can be rewritten in the matrix form as

$$\begin{aligned} & \frac{1}{2}\left(\mathbf{I}_{P^2} \otimes vec\left[\widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right) \frac{\widehat{b}_D^{(1)\,\prime}}{\widehat{b}_D} \widehat{h}^{(-2)}\right]'\right) vec\left(\mathbf{K}_{PP} \widehat{\mathbf{f}}^{(2)}\right) \\ = \;\; & vec\left[\widehat{h}^{(-2)} \widehat{h}^{(3)\prime} vec\left(\widehat{h}^{(-2)}\right) \frac{\widehat{b}_D^{(1)\,\prime}}{\widehat{b}_D} \widehat{h}^{(-2)}\right]. \end{aligned}$$

We can also get

$$B_3^2 = B_1^2 \times B_4^1 = \left(vec\left(\widehat{h}^{(-2)}\right) + \widehat{\mathbf{f}}^{(1)} B_1^1\right) B_4^1, \tag{48}$$

where

$$\widehat{\mathbf{f}}^{(1)} B_1^1 = vec\left(B_1^1 \widehat{\boldsymbol{\theta}}' + \widehat{\boldsymbol{\theta}} B_1^{1\prime}\right).$$

It is noted that

$$\begin{aligned} \overline{\boldsymbol{\theta}} \;\; & = \;\; \widehat{\boldsymbol{\theta}} + \frac{1}{n} B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_3^1\right) + O_p\left(\frac{1}{n^3}\right) \\ & = \;\; \widehat{\boldsymbol{\theta}} + \frac{1}{n} B_1^1 + \frac{1}{n^2}\left(B_2^1 - B_4^1 B_1^1\right) + O_p\left(\frac{1}{n^3}\right). \end{aligned}$$

Thus, we get

$$\begin{aligned} vec\left(\overline{\boldsymbol{\theta}}\overline{\boldsymbol{\theta}}'\right) \;\; = \;\; & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right) + \frac{1}{n} vec\left(\widehat{\boldsymbol{\theta}} B_1^{1\prime} + B_1^1 \widehat{\boldsymbol{\theta}}'\right) \\ & + \frac{1}{n^2} vec\left[\widehat{\boldsymbol{\theta}}\left(B_2^1 - B_4^1 B_1^1\right)' + \left(B_2^1 - B_4^1 B_1^1\right)\widehat{\boldsymbol{\theta}}'\right] + O_p\left(\frac{1}{n^3}\right). \end{aligned}$$

From (44), (45) and (48), we can show that

$$
\begin{aligned}
&\frac{\int vec\left(\boldsymbol{\theta}\boldsymbol{\theta}'\right) b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta} \\
= \quad & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right)+\frac{1}{n}B_1^2+\frac{1}{n^2}\left(B_2^2-B_3^2\right)+O_p\left(\frac{1}{n^3}\right) \\
= \quad & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right)+\frac{1}{n}\left[vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}B_1^1\right]+\frac{1}{n^2}\left(B_{21}^2+B_{22}^2-B_3^2\right)+O_p\left(\frac{1}{n^3}\right) \\
= \quad & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right)+\frac{1}{n}\left[vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}B_1^1\right]+\frac{1}{n^2}\left(B_{21}^2+B_{22}^2-B_3^2\right)+O_p\left(\frac{1}{n^3}\right) \\
= \quad & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right)+\frac{1}{n}\left[vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}B_1^1\right] \\
& +\frac{1}{n^2}\left[\widehat{\mathbf{f}}^{(1)}B_2^1+B_{22}^2-B_4^1\left(vec\left(\widehat{h}^{(-2)}\right)+\widehat{\mathbf{f}}^{(1)}B_1^1\right)\right]+O_p\left(\frac{1}{n^3}\right) \\
= \quad & vec\left(\widehat{\boldsymbol{\theta}}\widehat{\boldsymbol{\theta}}'\right)+\frac{1}{n}\left[vec\left(\widehat{h}^{(-2)}\right)+vec\left(B_1^1\widehat{\boldsymbol{\theta}}'+\widehat{\boldsymbol{\theta}}B_1^{1\prime}\right)\right] \\
& +\frac{1}{n^2}\left[vec\left(B_2^1\widehat{\boldsymbol{\theta}}'+\widehat{\boldsymbol{\theta}}B_2^{1\prime}\right)+B_{22}^2-B_4^1\left(vec\left(\widehat{h}^{(-2)}\right)+vec\left(B_1^1\widehat{\boldsymbol{\theta}}'+\widehat{\boldsymbol{\theta}}B_1^{1\prime}\right)\right)\right]+O_p\left(\frac{1}{n^3}\right).
\end{aligned}
$$

Hence we have

$$
\begin{aligned}
&\frac{\int vec\left[\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right)'\right] b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta} \\
= \quad & \frac{\int vec\left(\boldsymbol{\theta}\boldsymbol{\theta}'\right) b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}{\int b_D\left(\theta\right)\exp\left(-nh_n\left(\theta\right)\right)d\theta}-vec\left(\overline{\boldsymbol{\theta}\boldsymbol{\theta}}\right) \\
= \quad & \frac{1}{n}vec\left(\widehat{h}^{(-2)}\right)+\frac{1}{n^2}\left[B_{22}^2-B_4^1 vec\left(\widehat{h}^{(-2)}\right)\right]+O_p\left(\frac{1}{n^3}\right).
\end{aligned}
$$

We can further decompose $B_{22}^2-B_4^1 vec\left(\widehat{h}^{(-2)}\right)$ as

$$
B_{22}^2-B_4^1 vec\left(\widehat{h}^{(-2)}\right)=F_1+F_2,
$$

where

$$
\begin{aligned}
F_1 \quad = \quad & -\frac{5}{16}vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\left(\widehat{h}^{(-2)}\otimes vec\left(\widehat{h}^{(-2)}\right)'\right)\widehat{h}^{(4)}\right] \\
& +\frac{35}{48}vec\left(\widehat{h}^{(-2)}\right)\left[vec\left(\widehat{h}^{(-2)}\right)'\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\right] \\
& -vec\left(\widehat{h}^{(-2)}\right)\left(\frac{5}{24}\left[vec\left(\widehat{h}^{(-2)}\right)'\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\right]+\frac{1}{8}\mathbf{tr}\left[\left[\widehat{h}^{(-2)}\otimes vec\left(\widehat{h}^{(-2)}\right)\right]\widehat{h}^{(4)\prime}\right]\right) \\
= \quad & -\frac{7}{16}vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\left(\widehat{h}^{(-2)}\otimes vec\left(\widehat{h}^{(-2)}\right)'\right)\widehat{h}^{(4)}\right] \\
& +\frac{25}{48}vec\left(\widehat{h}^{(-2)}\right)\left[vec\left(\widehat{h}^{(-2)}\right)'\widehat{h}^{(3)}\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\right],
\end{aligned}
$$

and

$$\begin{aligned}
F_2 &= -\frac{5}{2}vec\left[\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\frac{\widehat{b}_D^{(1)\prime}}{\widehat{b}_D}\widehat{h}^{(-2)}\right]+\frac{3}{4}vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}^{(2)}}{\widehat{b}_D}\right]\\
&\quad -vec\left(\widehat{h}^{(-2)}\right)\left(\frac{1}{2}\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(2)}}{\widehat{b}_D}\right]-\frac{1}{2}vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}\right)\\
&= -\frac{5}{2}vec\left[\widehat{h}^{(-2)}\widehat{h}^{(3)\prime}vec\left(\widehat{h}^{(-2)}\right)\frac{\widehat{b}_D^{(1)\prime}}{\widehat{b}_D}\widehat{h}^{(-2)}\right]+\frac{1}{4}vec\left(\widehat{h}^{(-2)}\right)\mathbf{tr}\left[\widehat{h}^{(-2)}\frac{\widehat{b}^{(2)}}{\widehat{b}_D}\right]\\
&\quad +\frac{1}{2}vec\left(\widehat{h}^{(-2)}\right)vec\left(\widehat{h}^{(-2)}\right)^{\prime}\widehat{h}^{(3)}\widehat{h}^{(-2)}\frac{\widehat{b}_D^{(1)}}{\widehat{b}_D}.
\end{aligned}$$

This is the end of proof for this lemma. ■

6.2 Proof of Theorem 4.1

It is noted that $h_n(\boldsymbol{\theta}) = -\bar{l}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n} l_t(\boldsymbol{\theta})$, $b_D(\boldsymbol{\theta}) = p(\boldsymbol{\theta})$, $\pi(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta})$ and $\bar{\mathbf{H}}_n^{(j)}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} l_t^{(j)}(\boldsymbol{\theta}) = \bar{l}_n^{(j)}(\boldsymbol{\theta})$ for $j = 3, 4,$. Thus, according to Lemma 4.1, we have

$$\begin{aligned}
\overline{\boldsymbol{\theta}} &= \frac{\int \boldsymbol{\theta} p(\boldsymbol{\theta})\exp(-nh_n(\boldsymbol{\theta}))\,d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})\exp(-nh_n(\boldsymbol{\theta}))\,d\boldsymbol{\theta}} = \widehat{\boldsymbol{\theta}} - \frac{1}{n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}\\
&\quad +\frac{1}{2n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)^{\prime}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)+O_p\left(\frac{1}{n^2}\right), \qquad (49)
\end{aligned}$$

and

$$\begin{aligned}
vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) &= \frac{\int vec\left[\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}\right)^{\prime}\right]p(\boldsymbol{\theta})\exp(-nh_n(\boldsymbol{\theta}))\,d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta})\exp(-nh_n(\boldsymbol{\theta}))\,d\boldsymbol{\theta}}\\
&= -\frac{1}{n}vec\left(\hat{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)+\frac{1}{n^2}F_1+\frac{1}{n^2}F_2+O_p\left(\frac{1}{n^3}\right), \qquad (50)
\end{aligned}$$

where

$$\begin{aligned}
F_1 &= -\frac{7}{16}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)^{\prime}\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right]\\
&\quad +\frac{25}{48}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\left[vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)^{\prime}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)^{\prime}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right]\\
F_2 &= -\frac{5}{2}vec\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)^{\prime}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right]\\
&\quad +\frac{1}{4}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right]\\
&\quad +\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)^{\prime}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}.
\end{aligned}$$

From (49), by the Taylor expansion of $vec\left(\bar{\mathbf{H}}_n\left(\overline{\boldsymbol{\theta}}\right)\right)$ at $\widehat{\boldsymbol{\theta}}$, we have

$$vec\left(\bar{\mathbf{H}}_n\left(\overline{\boldsymbol{\theta}}\right)\right) = vec\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right) + \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right] + O_p\left(\frac{1}{n^2}\right).$$

Hence, we get

$$\begin{aligned} P_D^I &= \mathbf{tr}\left[-n\bar{\mathbf{H}}_n\left(\overline{\boldsymbol{\theta}}\right)V\left(\overline{\boldsymbol{\theta}}\right)\right] = -nvec\left(\bar{\mathbf{H}}_n\left(\overline{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\ &= -nvec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) - nvec\left(\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\ &\quad -nvec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) O_p\left(\frac{1}{n^2}\right) \end{aligned} \tag{51}$$

By (42), (43), and (44), we can have

$$\begin{aligned} & nvec\left(\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\ = \;& vec\left(\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right)' vec\left(nV\left(\overline{\boldsymbol{\theta}}\right)\right) \\ = \;& vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right]' \left[vec\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) + O_p(\frac{1}{n})\right] \\ = \;& vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right]' vec\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) + O_p(1)O_p(\frac{1}{n})O_p(\frac{1}{n}) \\ = \;& vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(\overline{\boldsymbol{\theta}} - \widehat{\boldsymbol{\theta}}\right)\right]' vec\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) + O_p(\frac{1}{n^2}) \\ = \;& vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{1}{n}\frac{\widehat{p}^{(1)}}{\widehat{p}} + \frac{1}{2n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) + O_p(\frac{1}{n^2})\right)\right]' \\ & \left[vec\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right] + O_p(\frac{1}{n^2}) \\ = \;& \frac{1}{n}vec\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} + \frac{1}{2}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right)\right]' \\ & \left[vec\left(-\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right] + O_p(\frac{1}{n^2}) \end{aligned}$$

$$\begin{aligned} & nvec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\ = \;& vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' nvec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\ = \;& vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' \left[-vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) + \frac{1}{n}F_1 + \frac{1}{n}F_2\right] + O_p\left(\frac{1}{n^2}\right) \\ = \;& -P + \frac{1}{n}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' F_1 + \frac{1}{n}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' F_2 + O_p\left(\frac{1}{n^2}\right). \end{aligned} \tag{52}$$

Furthermore, it can be shown that

$$\begin{aligned}
& vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' F_1 \\
= \; & -\frac{7}{16}P\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right] \\
& +\frac{25}{48}P\left[vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right], \qquad (53)
\end{aligned}$$

$$\begin{aligned}
& vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' F_2 \\
= \; & -\frac{5}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' vec\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right] \\
& +\frac{1}{4}P\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right]+\frac{1}{2}Pvec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} \qquad (54)
\end{aligned}$$

where

$$\begin{aligned}
& -\frac{5}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' vec\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right] \\
= \; & -\frac{5}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right] \\
= \; & -\frac{5}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right] \\
= \; & -\frac{5}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}. \qquad (55)
\end{aligned}$$

Hence, from (54) and (55), it is easy to show that

$$\begin{aligned}
vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' F_2 \; = & \; -\frac{5-P}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} \\
+ & \; \frac{P}{4}\mathbf{tr}\left[\hat{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right]. \qquad (56)
\end{aligned}$$

And from (52), (53) and (56), we can get that

$$\begin{aligned}
& nvec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right)' vec\left(V\left(\overline{\boldsymbol{\theta}}\right)\right) \\
= \; & -P-\frac{7}{16}P\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right] \\
& +\frac{25}{48}P\left[vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right] \\
& -\frac{5-P}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}}+\frac{1}{4}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] \qquad (57)
\end{aligned}$$

Then, from (51), (52) and (57), we have

$$P_D^I = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(\frac{1}{n^2}\right),$$

where

$$\begin{aligned}
C_1 &= \frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) \\
&\quad + \frac{7}{16}P\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right] \\
&\quad - \frac{25}{48}P\left[vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right] \\
&= \frac{7}{16}P\mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right] \\
&\quad + \frac{24-25P}{48}P\left[vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)\right],
\end{aligned}$$

$$\begin{aligned}
C_2 &= -\frac{\widehat{p}^{(1)\prime}}{\widehat{p}}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right) \\
&\quad \frac{5-P}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} - \frac{P}{4}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] \\
&= \frac{3-P}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(1)}}{\widehat{p}} - \frac{P}{4}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\frac{\widehat{p}^{(2)}}{\widehat{p}}\right] \\
&= \frac{3-P}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right) \\
&\quad - \frac{P}{4}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\right] - \frac{P}{4}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right)'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right).
\end{aligned}$$

We can rewrite C_1 and C_2 as

$$C_1 = \frac{7P}{16}C_{11} + \frac{24-25P}{48}C_{12},$$

$$C_2 = \frac{3-P}{2}C_{21} - \frac{P}{4}C_{22} - \frac{P}{4}C_{23},$$

where

$$C_{11} = \mathbf{tr}\left[\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\right)\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}\right)\right],$$

$$C_{12} = vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right),$$

$$C_{21} = vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right),$$

$$C_{22} = \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(2)}\left(\widehat{\boldsymbol{\theta}}\right)\right],\ C_{23} = \pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right)'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}\right)^{-1}\pi^{(1)}\left(\widehat{\boldsymbol{\theta}}\right).$$

And from Li et al (2017)

$$\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) = \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right) - \frac{1}{2n}C_{21} + \frac{1}{2n}C_{23} + \frac{1}{8n}C_{12} + O_p\left(n^{-2}\right).$$

Hence

$$\begin{aligned}
\text{IDIC} &= -2\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) + 2P_D^I \\
&= -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right) + \frac{1}{n}C_{21} - \frac{1}{n}C_{23} - \frac{1}{4n}C_{12} + 2P + \frac{2}{n}C_1 + \frac{2}{n}C_2 + O_p\left(\frac{1}{n^2}\right) \\
&= -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}\right) + 2P + \frac{1}{n}\left(-\frac{C_{12}}{4} + \frac{7P}{8}C_{11} + \frac{24-35P}{24}C_{12}\right) \\
&\quad + \frac{1}{n}\left(\left(3-P\right)C_{21} - \frac{P}{2}C_{22} - \frac{P}{2}C_{23} + C_{21} - C_{23}\right) + O_p\left(\frac{1}{n^2}\right) \\
&= \text{AIC} + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(\frac{1}{n^2}\right),
\end{aligned}$$

where

$$D_1 = \frac{7P}{8}C_{11} + \frac{18-25P}{24}C_{12},$$

$$D_2 = \left(4-P\right)C_{21} - \frac{P}{2}C_{22} - \frac{2+P}{2}C_{23}.$$

References

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

Avramov, D. and Zhou, G.F. (2010). Bayesian portfolio analysis. *Annual Review of Financial Economics*, **2**, 25-47.

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.

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.

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

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

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

Chan, J. C., and Grant, A. L. (2016a). On the observed-data deviance information criterion for volatility modeling. *Journal of Financial Econometrics*, **14(4)**, 772-802.

Chan, J. C., and Grant, A. L. (2016b). Fast computation of the deviance information criterion for latent variable models. *Computational Statistics and Data Analysis*, **100**, 847-859.

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

Creal, D. (2012). A survey of sequential Monte Carlo methods for economics and finance. *Econometric reviews*, **31(3)**, 245-296.

Doucet, A., and Johansen, A.M. (2009). A tutorial on particle filtering and smoothing: Fifteen years later. *Handbook of Nonlinear Filtering*, **12**, 656-704.

Doucet, A., and Shephard, N. (2012). Robust inference on parameters via particle filters and sandwich covariance matrices. Working Paper, University of Oxford, Department of Economics.

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.

Fama E. F., and K. R. French. (1993). Common risk factors in the returns on stocks and bonds. *Journal of Financial Economics*, **33**, 3-56.

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

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.

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

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

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

Herbst, E. (2010). Gradient and Hessian-based MCMC for DSGE Models. Working Paper.

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.

Iskrev, N. (2008). Evaluating the information matrix in linearized DSGE models. *Economics Letters*, **99(3)**, 607-610.

Isserlis, L. (1918) On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, **12**, 134–139.

Kan, R. and Zhou, G.F. (2006). Modeling Non-normality using multivariate t: implications for asset pricing. Working Paper.

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.

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

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

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

Li, Y., Yu, J., and Zeng, T. (2017). Deviation information criterion for Bayesian model comparison: justification and variation, Working Paper, Singapore Management University.

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

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

Millar, R. B. (2009). Comparison of hierarchical Bayesian models for over-dispersed count data using DIC and Bayes factors. *Biometrics*. **65**, 962-969.

Millar, R. B, and S. McKechnie. (2014). A one-step-ahead pseudo DIC for comparison of Bayesian state-space models. *Biometrics*. **70**, 972-980.

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

Norets, A. (2009). Inference in dynamic discrete choice models with serially correlated unobserved state variables. *Econometrica*, **77(5)**, 1665-1682.

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

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

Poyiadjis, G., Doucet, A., Singh, S. S. (2011). Particle approximations of the score and observed information matrix in state-space models with application to parameter estimation. *Biometrika*, **98(1)**, 65-80.

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.

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.

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.*

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.

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.

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

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

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

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.

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

Vignat, C. (2012). A generalized Isserlis theorem for location mixtures of Gaussian random vectors. *Statistics & Probability Letters*, **82(1)**, 67-71.

White, H. (1996). Estimation, inference and specification analysis. *Cambridge university press.*

Withers, C.S., (1985) The moments of the multivariate normal, *Bulletin of the Australian Mathematical Society*, **32**, 103-107

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

Zhou, G.F. (1993). Asset pricing testing under alternative distributions. *Journal of Finance*, **5**, 1927-1942.