2-2017

Deviance information criterion for Bayesian model selection: Justification and variation

Yong LI
Renmin University of China

Jun YU
Singapore Management University, yujun@smu.edu.sg

Tao ZENG
Zhejiang University

SMU ECONOMICS & STATISTICS

Deviance Information Criterion for Bayesian Model Selection: Justification and Variation

Yong Li, Jun Yu and Tao Zeng

February 2017

Paper No. 05-2017

Deviance Information Criterion for Bayesian Model Selection: Justification and Variation*

Yong Li
Renmin University of China

Jun Yu
Singapore Management University

Tao Zeng
Wuhan University

February 15, 2017

Abstract

Deviance information criterion (DIC) has been extensively used for making Bayesian model selection. It is a Bayesian version of AIC and chooses a model that gives the smallest expected Kullback-Leibler divergence between the data generating process (DGP) and a predictive distribution asymptotically. We show that when the plug-in predictive distribution is used, DIC can have a rigorous decision-theoretic justification under regularity conditions. An alternative expression for DIC, based on the Bayesian predictive distribution, is proposed. The new DIC has a smaller penalty term than the original DIC and is very easy to compute from the MCMC output. It is invariant to reparameterization and yields a smaller frequentist risk than the original DIC asymptotically.

JEL classification: C11, C12, G12
Keywords: AIC; DIC; Bayesian Predictive Distribution; Plug-in Predictive Distribution; Loss Function; Bayesian Model Comparison; Frequentist Risk

1 Introduction

A highly important statistical inference often faced by model builders and empirical researchers is model selection (Phillips, 1995, 1996). Many penalty-based information criteria have been proposed to select from candidate models. In the frequentist statistical framework, the most popular information criteria are AIC and BIC. Arguably one of the most important

*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, Economics and Management School, Wuhan University, Wuhan, China 430072. Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No. 71271221), Program for New Century Excellent Talents in University.

developments in the Bayesian literature in recent years is the deviance information criterion (DIC) of Spiegelhalter, et al (2002) for model selection.[1] DIC is a Bayesian version of AIC. Like AIC, it trades off a measure of model adequacy against a measure of complexity and is concerned with how replicate data predict the observed data. Unlike AIC, DIC takes prior information into account.

DIC is constructed based on the posterior distribution of the log-likelihood or the deviance, and has several desirable features. Firstly, DIC is easy to calculate when the likelihood function is available in closed-form and the posterior distributions of the models are obtained by Markov chain Monte Carlo (MCMC) simulation. Secondly, it is applicable to a wide range of statistical models. Thirdly, unlike the Bayes factors (BF), it is not subject to the Jeffreys-Lindley's paradox and can be calculated when noninformative or improper priors are used.

However, as acknowledged in Spiegelhalter, et al (2002, 2014), the decision-theoretic justification of DIC is not rigorous in the literature and DIC is not invariant to reparameterization. The first contribution of the present paper is to provide a rigorous decision-theoretic justification to DIC when the standard Bayesian large sample theory is valid and when the data are not necessarily independent. It can be shown that DIC is an asymptotically unbiased estimator of the expected Kullback-Leibler (KL) divergence between the data generating process (DGP) and the plug-in predictive distribution, when the Bayesian estimate is used. This justification is the same as how AIC has been justified.

In the Bayesian framework, an alterative predictive distribution to the plug-in predictive distribution is the Bayesian predictive distribution. Naturally, the KL divergence between the DGP and the Bayesian predictive distribution can be used as the loss function which can in turn be used to derive a new information criterion for model comparison. Unlike the plug-in predictive distribution, the Bayesian predictive distribution is invariant to reparameterization. Recently Ando and Tsay (2010) developed an information criterion that provides asymptotically unbiased estimation to the new expected KL divergence in the independent and identically distributed (iid) environment. The second contribution of the present paper is to develop a new information criterion that provides an asymptotically unbiased estimation to the new expected KL divergence under a general framework. Relaxing the iid assumption is important because the iid assumption is often violated in practice. Moreover, compared with the information criterion developed in Ando and Tsay (2010), our information criterion has a simpler expression. It is easier to compare our information criterion with other information criteria. Furthermore, it is trivial to compute from DIC.

Our theoretical results shows that asymptotically the frequentist risk implied by the

[1] According to, Spiegelhalter et al. (2014), Spiegelhalter et al. (2002) was the third most cited paper in international mathematical sciences between 1998 and 2008. Up to January 2017, it has received 4318 citations on the Web of Knowledge and over 7587 on Google Scholar.

Bayesian predictive distribution is smaller than that implied by the plug-in predictive distribution. Hence, from the predictive viewpoint, the Bayesian predictive distribution is a better predictive distribution. This represents another important advantage of using the Bayesian predictive distribution and hence our new information criterion.

The paper is organized as follows. Section 2 explains how to treat the model selection as a decision problem and gives a simple review about the decision-theoretic justification of AIC. Section 3 provides a rigorous decision-theoretic justification to DIC of Spiegelhalter, et al (2002) under a set of regularity conditions, and shows that why DIC can be explained as Bayesian version of AIC. In Section 4, based on the Bayesian predictive distribution, a new information criterion is proposed. Its theoretical properties are established and comparisons with other information criteria are also made in this section. Section 5 concludes the paper. The Appendix collects the proof of the theoretical results in the paper.

2 Decision-theoretic Justification of AIC

There are essentially two strands of literature on model selection.[2] The first strand aims to answer the following question – which model best explains the observed data? The BF (Kass and Raftery, 1995) and its variations belong to this strand. They compare models by examining "posterior probabilities" given the observed data and search for the "true" model. BIC is a large sample approximation to BF although it is based on the likelihood function. The second strand aims to answer the following question – which model give the best predictions of future observations generated by the same mechanism that gives rise to the observed data? Clearly this is a utility-based approach where the utility is set to be the prediction. Ideally, we would like to choose the model that gives the best overall predictions of future values. Some cross validation-based criteria have been developed where the original sample into a training and a validation set (Vehtari and Lampinen, 2002; Zhang and Yang, 2015). Unfortunately, different ways of sample splitting often lead to different outcomes. Alternatively, based on hypothetically replicate data generated by the same mechanism that gives rise to the observed data, some predictive information criteria have been proposed for model selection. They minimize a loss function associated with the predictive decisions. AIC and DIC are two well-known criteria in this framework. After the decision is made about which model should be used for prediction, a unique prediction action for future observations can be obtained to fulfill the original goal. This last approach is what we follow in the present paper.

[2]For more information about the literature, see Vehtari and Ojanen (2012) and Burnham and Anderson (2002).

2.1 Predictive model selection as a decision problem

Assuming that the probabilistic behavior of observed data, $\mathbf{y} \in \mathbf{Y}$, is described by a set of probabilistic models such as $\{M_k\}_{k=1}^K := \{p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)\}_{k=1}^K$ where parameter $\boldsymbol{\theta}_k$ is the set of parameters in model M_k. Formally, the model selection problem can be taken as a decision problem to select a model among $\{M_k\}_{k=1}^K$ where the action space has K elements, namely, $\{d_k\}_{k=1}^K$, where d_k means M_k is selected.

For the decision problem, a loss function, $\mathcal{L}(\mathbf{y}, d_k)$, which measures the loss of decision d_k as a function of $\mathbf{y}$, must be specified. Given the loss function, the frequentist risk can be defined as (Berger, 1985)

$$Risk(d_k) = E_{\mathbf{y}}\left[\mathcal{L}(\mathbf{y}, d_k)\right] = \int \mathcal{L}(\mathbf{y}, d_k) g(\mathbf{y}) \mathrm{d}\mathbf{y},$$

where $g(\mathbf{y})$ is the DGP of $\mathbf{y}$. Hence, the model selection problem is equivalent to optimizing the statistical decision,

$$k^* = \arg\min_k Risk(d_k).$$

Based on the set of candidate models $\{M_k\}_{k=1}^K$, the model M_{k^*} with the decision d_{k^*} is selected.

Let $\mathbf{y}_{rep}$ be the replicate data independently generated by the same mechanism that gives rise to the observed data $\mathbf{y}$. Assume the sample size in $\mathbf{y}_{rep}$ is the same as that in $\mathbf{y}$. Consider the predictive density of this replicate experiment for a candidate model M_k. The plug-in predictive density can be expressed as $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)$ for M_k where $\widehat{\boldsymbol{\theta}}_k(\mathbf{y})$ is the quasi maximum likelihood (QML) estimate of $\boldsymbol{\theta}_k$, obtained from $\mathbf{y}$ and defined by

$$\widehat{\boldsymbol{\theta}}_k(\mathbf{y}) = \arg\max_{\boldsymbol{\theta}_k} \ln p(\mathbf{y}|\boldsymbol{\theta}_k, M_k).$$

The quantity that has been used to measure the quality of the candidate model in terms of its ability to make predictions is the KL divergence between $g(\mathbf{y}_{rep})$ and $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)$ multiplied by 2,

$$2 \times KL\left[g(\mathbf{y}_{rep}), p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)\right] = 2E_{\mathbf{y}_{rep}}\left[\ln \frac{g(\mathbf{y}_{rep})}{p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)}\right]$$

$$= 2\int \left[\ln \frac{g(\mathbf{y}_{rep})}{p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)}\right] g(\mathbf{y}_{rep}) \mathrm{d}\mathbf{y}_{rep}.$$

Naturally the loss function associated with decision d_k is

$$\mathcal{L}(\mathbf{y}, d_k) = 2 \times KL\left[g(\mathbf{y}_{rep}), p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)\right].$$

As a result, the model selection problem is,

$$\begin{aligned} k^* &= \arg\min_k Risk(d_k) = \arg\min_k E_{\mathbf{y}}\left[\mathcal{L}(\mathbf{y}, d_k)\right] \\ &= \arg\min_k \left\{ 2 \times E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)} \right] \right\} \\ &= \arg\min_k \left\{ E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[2 \ln g\left(\mathbf{y}_{rep}\right)\right] + E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right) \right] \right\}. \end{aligned}$$

Since $g\left(\mathbf{y}_{rep}\right)$ is the DGP, $E_{\mathbf{y}_{rep}}\left[2 \ln g\left(\mathbf{y}_{rep}\right)\right]$ is the same across all candidate models, and, hence, is dropped from the above equation. Consequently, we have

$$k^* = \arg\min_k Risk(d_k) = \arg\min_k E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right) \right].$$

The smaller the $Risk(d_k)$, the better the candidate model performs when using $p(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k)$ to predict $g\left(\mathbf{y}_{rep}\right)$. The optimal decision makes it necessary to evaluate the risk. AIC provides an asymptotically unbiased estimation of $E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right) \right]$.

2.2 AIC for predictive model selection

To show that AIC provides an unbiased estimation of $E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right) \right]$ asymptotically, let us first fix some notations. When there is no confusion, we simply write candidate model $p\left(\mathbf{y}|\boldsymbol{\theta}_k, M_k\right)$ as $p(\mathbf{y}|\boldsymbol{\theta})$ where $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_P)' \in \boldsymbol{\Theta} \subseteq R^P$. Under the iid assumption, let $\mathbf{y} = (y_1, y_2, \cdots, y_n)'$ denote the observed data, $\mathbf{y}_{rep} = (y_{1,rep}, \cdots, y_{n,rep})'$ denote the replicate data, and n be the sample size in both sets of data. Although $-2 \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right)$ is a natural estimate of $E_{\mathbf{y}_{rep}}\left(-2 \ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right)$, it is asymptotically biased because $\mathbf{y}$ has been used twice. Let

$$c(\mathbf{y}) = E_{\mathbf{y}_{rep}}\left(-2 \ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right)\right) - \left(-2 \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right)\right). \tag{1}$$

Under a set of regularity conditions, one can show that $E_{\mathbf{y}}\left(c(\mathbf{y})\right) \to 2P$. Hence, if we let AIC $= -2 \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}(\mathbf{y})\right) + 2P$, then, as $n \to \infty$,

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

To see why a penalty term, $2P$, is needed in AIC, let

$$\boldsymbol{\theta}^t := \arg\min_{\theta} \frac{1}{n} KL[g(\mathbf{y}), p(\mathbf{y}|\boldsymbol{\theta})] \tag{2}$$

be the pseudo-true parameter value; $\widehat{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)$ be the QML estimate of $\boldsymbol{\theta}$ obtained from $\mathbf{y}_{rep}$;

$p(\mathbf{y}|\boldsymbol{\theta})$ be a "good approximation" to the DGP. Note that

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

Clearly, the term in $T1$ is the same as $E_{\mathbf{y}} \left(-2 \ln p \left(\mathbf{y} | \widehat{\boldsymbol{\theta}}(\mathbf{y}) \right) \right)$. The term in $T2$ is the expectation of the likelihood ratio statistic based on the replicate data. Under a set of regularity conditions that ensure $\sqrt{n}$-consistency and asymptotic normality of the QML estimate, we have $T2 = T3 + o(1)$. To approximate the term in $T3$, if $\widehat{\boldsymbol{\theta}}(\mathbf{y})$ is a consistent estimate of $\boldsymbol{\theta}^t$, we have

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

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

$$E_{\mathbf{y}} \left\{ -2 E_{\mathbf{y}_{rep}} \left[\frac{\partial \ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t \right)'}{\partial \theta} \left(\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^t \right) \right] \right\} = -2 E_{\mathbf{y}_{rep}} \left(\frac{\partial \ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t \right)}{\partial \theta} \right)' E_{\mathbf{y}} \left(\widehat{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^t \right) = 0.$$

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

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

where $\mathbf{tr}$ denotes the trace of a matrix. Following Burnham and Anderson (2002), we have

$$E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}(\mathbf{y}) \right) \right) = E_{\mathbf{y}} \left(-2 \ln p \left(\mathbf{y} | \widehat{\boldsymbol{\theta}}(\mathbf{y}) \right) + 2P \right) + o(1) = E_{\mathbf{y}} \left(\text{AIC} \right) + o(1),$$

that is, AIC is an unbiased estimator of $E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \widehat{\boldsymbol{\theta}}(\mathbf{y}) \right) \right)$ asymptotically. From the decision viewpoint, among candidate models, AIC selects a model which minimizes the frequentist risk when the plug-in predictive distribution is used for making predictions.

It is clear that the decision-theoretic justification of AIC requires a careful choice of the KL divergence function, the use of QML estimation, and a set of regularity conditions that ensure

the $\sqrt{n}$-consistency and the asymptotic normality of the QML estimates. The penalty term in AIC arises from two sources. First, the pseudo-true value has to be estimated. Second, the estimate obtained from the observed data is not the same as that from the replicate data. Moreover, as pointed out in Burnham and Anderson (2002), the justification of AIC requires the candidate model be a "good approximation" to the DGP for the trace to be P asymptotically in $T3$. However, Burnham and Anderson (2002) did not provide the formal definition of "good approximation".

3 Decision-theoretic Justification of DIC

3.1 DIC

Spiegelhalter, et al (2002) proposed DIC for Bayesian model selection. The criterion is based on the deviance

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

and takes the form of

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

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

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

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

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

where $\bar{\boldsymbol{\theta}}(\mathbf{y})$ is the Bayesian estimator based on $\mathbf{y}$, and more precisely the posterior mean of $\boldsymbol{\theta}$, $\int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}$. When there is no confusion, we simple write $\bar{\boldsymbol{\theta}}(\mathbf{y})$ as $\bar{\boldsymbol{\theta}}$.

DIC can be rewritten by in another two equivalent forms:

$$\text{DIC} = D\left(\bar{\boldsymbol{\theta}}\right) + 2P_D, \tag{5}$$

and

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

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

However, as acknowledged in Spiegelhalter et al. (2002) (Section 7.3 on Page 603 and the first paragraph on Page 605), the justification of DIC is informal and heuristic. In this section, we provide a rigorous decision-theoretic justification of DIC, in the same spirit as the justification of AIC. We show that, when a proper loss function is selected, DIC is an asymptotically unbiased estimator of the loss function.

3.2 Decision-theoretic justification of DIC

When developing DIC, Spiegelhalter, et al (2002) did not explicitly specify the KL divergence function. However, from Equation (33) on Page 602, the loss function defined in the first paragraph on Page 603, and Equation (40) on Page 603 in their paper, one may deduce that the following KL divergence[3]

$$KL\left[p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] = E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left[\ln \frac{p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)}{p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)}\right] \tag{7}$$

was used. Hence,

$$2 \times KL\left[p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] = 2 \times E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right) + E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right). \tag{8}$$

In Equation (33), Spiegelhalter, et al dealt with $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right)$ only and ignored the first term in the right hand side of Equation (8). On Page 604, they argued that, if

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

then

$$\int \left\{E_{\boldsymbol{\theta}|\mathbf{y}}\left[c\left(\mathbf{y}, \boldsymbol{\theta}, \bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right] - 2P_D\right\} p(\mathbf{y}) d\mathbf{y} \to \mathbf{0}, \tag{9}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})\mathrm{d}\boldsymbol{\theta}$. This leads to DIC $= D(\bar{\boldsymbol{\theta}}) + 2P_D$. The convergence in (9) was proved without specifying any conditions. Most importantly, an implicit assumption made in this heuristic argument is that the first term in the right hand side of Equation (8) is constant across candidate models and thus dropped from (8). While the treatment mimics the development of AIC, unfortunately, one cannot claim that $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right)$ is the

[3]In Equation (33) of Spiegelhalter, et al (2002), the expectation is taken with respect to $\mathbf{y}_{rep}|\boldsymbol{\theta}^t$ which corresponds to the candidate model. In AIC, the expectation is taken with respect to $\mathbf{y}_{rep}$ which corresponds to the DGP.

same across all candidate models. This is because, as Spiegelhalter, et al. (2002) stated in the second paragraph on Page 604, "we are taking a Bayesian perspective" and "we replace the pseudo-true value by a random quantity". As a result, $\boldsymbol{\theta}$ in the first term in the right hand side of Equation (8) is model dependent and in general $E_{\mathbf{y}_{rep}|\boldsymbol{\theta}}\left(\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right)$ takes a different value for each candidate model. Furthermore, as in AIC, the candidate model is required to be a "good approximation" to the DGP. However, as in Burnham and Anderson (2002), Spiegelhalter et al. (2002) did not provide the formal definition of "good approximation".

From the discussion above, clearly $KL\left[p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right), p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ is not a proper KL divergence function to justify DIC. A new KL divergence is needed. As in AIC, we first consider the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$ in the following KL divergence

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

The corresponding frequentist risk function of a statistical decision d_k for model selection is

$$\begin{aligned} Risk(d_k) &= E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left[2\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)}\right]\right\} \\ = \;& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[2\ln g\left(\mathbf{y}_{rep}\right)\right] + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)\right]. \end{aligned}$$

Since $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[2\ln g\left(\mathbf{y}_{rep}\right)\right]$ is the same across candidate models, minimizing the frequentist risk function $Risk(d_k)$ is equivalent to minimizing

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

Denote the selected model by M_{k^*}. Then $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_{k^*}(\mathbf{y}), M_{k^*}\right)$ is used to generate future observations where $\bar{\boldsymbol{\theta}}_{k^*}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}$ in M_{k^*}.

We are now in the position to provide a rigorous decision-theoretic justification of DIC based on a set of regularity conditions. Let $\mathbf{y}^t := (y_1, \ldots, y_t)$. Define $l_t\left(\boldsymbol{\theta}\right) = \ln p(\mathbf{y}^t|\boldsymbol{\theta}) - \ln p(\mathbf{y}^{t-1}|\boldsymbol{\theta})$ to be the conditional log-likelihood for the t^{th} observation and $\nabla^j l_t\left(\boldsymbol{\theta}\right)$ to be the j^{th} derivative of $l_t\left(\boldsymbol{\theta}\right)$, $\nabla^j l_t\left(\boldsymbol{\theta}\right) = l_t\left(\boldsymbol{\theta}\right)$ when $j = 0$. We suppress the superscript when $j = 1$, and

$$\begin{aligned} &\mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) := \frac{\partial \ln p(\mathbf{y}^t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{i=1}^{t} \nabla l_i\left(\boldsymbol{\theta}\right), \; \mathbf{h}(\mathbf{y}^t, \boldsymbol{\theta}) := \frac{\partial^2 \ln p(\mathbf{y}^t|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = \sum_{i=1}^{t} \nabla^2 l_i\left(\boldsymbol{\theta}\right), \\ &\mathbf{s}_t(\boldsymbol{\theta}) := \nabla l_t\left(\boldsymbol{\theta}\right) = \mathbf{s}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{s}(\mathbf{y}^{t-1}, \boldsymbol{\theta}), \; \mathbf{h}_t(\boldsymbol{\theta}) := \nabla^2 l_t\left(\boldsymbol{\theta}\right) = \mathbf{h}(\mathbf{y}^t, \boldsymbol{\theta}) - \mathbf{h}(\mathbf{y}^{t-1}, \boldsymbol{\theta}), \\ &\mathbf{B}_n\left(\boldsymbol{\theta}\right) := Var\left[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} \bigtriangledown l_t\left(\boldsymbol{\theta}\right)\right], \hat{\mathbf{H}}_n(\boldsymbol{\theta}) := \frac{1}{n} \sum_{t=1}^{n} \mathbf{h}_t(\boldsymbol{\theta}), \\ &L_n(\boldsymbol{\theta}) := \ln p(\boldsymbol{\theta}|\mathbf{y}), \; L_n^{(j)}(\boldsymbol{\theta}) := \partial^j \ln p(\boldsymbol{\theta}|\mathbf{y})/\partial \boldsymbol{\theta}^j. \end{aligned}$$

We further denote and $\mathbf{H}_n(\boldsymbol{\theta}) = \int \hat{\mathbf{H}}_n(\boldsymbol{\theta}) g\left(\mathbf{y}\right) d\mathbf{y}$.

In this paper, we impose the following regularity conditions.

Assumption 1: There exists a finite sample size n^* such that for $n > n^*$ there exists a local maximum of the posterior density $L_n(\boldsymbol{\theta})$ at $\overleftrightarrow{\boldsymbol{\theta}}$ that satisfies $L_n^{(1)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) = 0$ and $L_n^{(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right)$ is negative definite.

Assumption 2: The largest eigenvalue of $\left[-L_n^{(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right]^{-1}$ goes to zero in probability as $n \to \infty$.

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

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

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

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

$$\int_{\boldsymbol{\Theta} - H\left(\overleftrightarrow{\boldsymbol{\theta}}, \delta\right)} p\left(\boldsymbol{\theta}|\mathbf{y}\right) d\boldsymbol{\theta} = o_p\left(1\right),$$

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

Assumption 5: For any element of θ_i or θ_j, $i, j = 1, , \cdots, P$, we have

$$\int \theta_i^2 p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty, \int \theta_i^2 \theta_j^2 p(\boldsymbol{\theta}) \mathrm{d}\boldsymbol{\theta} < \infty.$$

Assumption 6: Let $\boldsymbol{\theta}^t \in \boldsymbol{\Theta} \subset R^P$ be the pseudo-true value that minimizes the KL loss between the DGP and the candidate model,

$$\boldsymbol{\theta}^t = \arg\min_{\boldsymbol{\theta}} \lim_{n \to \infty} \frac{1}{n} \int \ln \frac{g(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})} g(\mathbf{y}) d\mathbf{y},$$

where $\boldsymbol{\theta}^t$ is the unique minimizer.

Assumption 7: $\boldsymbol{\theta}^t \in int\left(\boldsymbol{\Theta}\right)$ where $\boldsymbol{\Theta}$ is a compact, separable metric space.

Assumption 8: $\{y_t\}_{t=1}^{\infty}$ is strong mixing with the mixing coefficient $\alpha\left(m\right) = O\left(m^{\frac{-2r}{r-2} - \varepsilon}\right)$ for some $\varepsilon > 0$ and $r > 2$.

Assumption 9: $l_t\left(\boldsymbol{\theta}\right)$ satisfies the standard measurability and the second order differentiability conditions on $\mathcal{F}_{-\infty}^t \times \boldsymbol{\Theta}$ where $\mathcal{F}_{-\infty}^t = \sigma\left(y_t, y_{t-1}, \cdots\right)$.

Assumption 10: There exists a function $M_t(\mathbf{y}^t)$ such that for $0 \leqslant j \leqslant 3$, all $\boldsymbol{\theta} \in \mathcal{G}$ where $\mathcal{G}$ is an open, convex set containing $\boldsymbol{\Theta}$, $\bigtriangledown^j l_t\left(\boldsymbol{\theta}\right)$ exists, $\sup_{\boldsymbol{\theta} \in \mathcal{G}} \left\|\bigtriangledown^j l_t\left(\boldsymbol{\theta}\right)\right\| \leqslant M_t(\mathbf{y}^t)$, and $\sup_t E\left\|M_t(\mathbf{y}^t)\right\|^{r+\delta} \leq M < \infty$ for some $\delta > 0$.

Assumption 11: $\left\{\bigtriangledown^j l_t\left(\boldsymbol{\theta}\right)\right\}$ 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 $j = 2, 3$ uniformly on $\boldsymbol{\Theta}$.

Assumption 12: $\lim_{n \to \infty} \frac{1}{n} \sum_{t=1}^{n} \int \bigtriangledown^j l_t\left(\boldsymbol{\theta}\right) g\left(\mathbf{y}\right) d\mathbf{y}$ exists for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $0 \leqslant j \leqslant 3$.

Assumption 13: For all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\left\| \bigtriangledown^j l_t(\boldsymbol{\theta}) - \bigtriangledown^j l_t(\boldsymbol{\theta}') \right\| \leq c_t(\mathbf{y}^t) \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$ for $0 \leqslant j \leqslant 3$ in probability, where $c_t(\mathbf{y}^t)$ is a positive random variable, $\sup_t E \left\| c_t(\mathbf{y}^t) \right\| < \infty$ and $\lim_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} (c_t - Ec_t) \xrightarrow{p} 0$.

Assumption 14: $\mathbf{H}(\boldsymbol{\theta}^t) = \lim_{n\to\infty} \mathbf{H}_n(\boldsymbol{\theta}^t)$ exists and is negative definite. $\mathbf{B}(\boldsymbol{\theta}^t) = \lim_{n\to\infty} \mathbf{B}_n(\boldsymbol{\theta}^t)$ exist and is positive definite.

Assumption 15: $\mathbf{H}_n(\boldsymbol{\theta}^t) = -\mathbf{B}_n(\boldsymbol{\theta}^t) + o(1)$.

Remark 3.1 *Assumptions 1-4 have been used in the literature to develop the standard Bayesian large sample theory; see, for example, Chen (1985), Kim (1994, 1998), Geweke (2005). Under Conditions 1-4, Chen (1985) shows that the posterior distribution converges to a normal distribution with the posterior mode being the mean and with the inverse of the second derivative of the log-posterior distribution evaluated at the mode being its covariance. Assumption 5 is used to ensure that the first and the second moments of the posterior distribution exist.*

Remark 3.2 *Assumptions 6-14 are the regularity conditions for the QML theory for dependent and heterogeneous data; see, for example, Andrews (1987, 1988), White (1996), Wooldridge (1994). If the data are iid, Assumption 6 is reduced to* $\boldsymbol{\theta}^t := \arg\min_\theta \frac{1}{n} KL[g(\mathbf{y}), p(\mathbf{y}|\boldsymbol{\theta})]$, *the definition in (2). For the iid data,* $E_{\mathbf{y}_{rep}}\left[\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)/\partial\boldsymbol{\theta}\right] = 0$. *Unfortunately, in general* $E_{\mathbf{y}_{rep}}\left[\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)/\partial\boldsymbol{\theta}\right] \neq 0$ *for dependent data because the conditional likelihood,* $l_t(\boldsymbol{\theta})$, *generally depends on* y^t, *the entire history of the observed data. This property renders the proof of asymptotic unbiasedness of DIC more challenging. That is why we need Assumption 11 to control the time dependence in* $l_t(\boldsymbol{\theta})$ *and its derivatives (Gallant and White, 1988).*

Remark 3.3 *Assumption 15 gives the exact definition of "good approximation". Combining with Assumption 14 that assumes* $\lim_{n\to\infty} \mathbf{H}_n(\boldsymbol{\theta}^t) = \mathbf{H}(\boldsymbol{\theta}^t)$ *and* $\mathbf{B}(\boldsymbol{\theta}^t) = \lim_{n\to\infty} \mathbf{B}_n(\boldsymbol{\theta}^t)$, *this condition entails the asymptotic validity of the information matrix identity, i.e.,* $\mathbf{H}(\boldsymbol{\theta}^t) = -\mathbf{B}(\boldsymbol{\theta}^t)$. *When the candidate model is correctly specified, the condition is satisfied, but not vice versa. In fact, even if* $\mathbf{H}_n(\boldsymbol{\theta}^t) = -\mathbf{B}_n(\boldsymbol{\theta}^t)$, *the model can be misspecified.*

Example 3.1 *Assumption 15 assumes that the information matrix identity holds asymptotically, but may not hold for any given sample size n. The following example to explain this point. Let the DGP be*

$$y_t = x_{1t}\beta_0 + x_{2t}\gamma_0 + \varepsilon_t,\ \ \varepsilon_t | x_{1t}, x_{2t} \overset{iid}{\sim} N(0, \sigma^2).$$

Following Claeskens and Hjort (2003), assume that $\gamma_0 = \delta_0/n^{1/2}$, *where* δ_0 *is an unknown constant. Let the candidate model be*

$$y_t = x_{1t}\beta + v_t, v_t | x_{1t} \overset{iid}{\sim} N(0, \sigma^2).$$

the quasi-likelihood function is

$$L_n\left(\mathbf{y}^n, \mathbf{x}_1^n|\boldsymbol{\theta}\right) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{n}\sum_{t=1}^{n}\frac{\left(y_t - x_{1t}\beta\right)^2}{2\sigma^2},$$

where $\boldsymbol{\theta} = \left(\beta', \sigma^2\right)'$, $\mathbf{y}^n = (y_1, ..., y_n)$ *and* $\mathbf{x}_1^n = (x_{11}, ..., x_{1n})$. *In this case, it is easy to show that the pseudo true value is* $\boldsymbol{\theta}^t = \left(\beta_0', \sigma_0^2\right)'$. *For the candidate model, the negative Hessian matrix is*

$$-\mathbf{H}_n\left(\boldsymbol{\theta}^t\right) = \frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} \frac{E\left(x_{1t}^2\right)}{\sigma_0^2} & \frac{E(x_{1t}x_{2t})\gamma_0}{\left(\sigma_0^2\right)^2} \\ \frac{\gamma_0 E(x_{1t}x_{2t})}{\left(\sigma_0^2\right)^2} & \frac{1}{2\left(\sigma_0^2\right)^2} + \frac{\gamma_0 E(x_{2t}x_{2t})\gamma_0}{\left(\sigma_0^2\right)^3} \end{bmatrix}$$

while the information matrix is

$$\begin{aligned} &\mathbf{B}_n\left(\boldsymbol{\theta}^t\right) \\ = \ & \frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} \frac{E\left(x_{1t}^2\right)}{\sigma_0^2} + \frac{E\left[(x_{2t}\gamma_0)^2 x_{1t}^2\right]}{\left(\sigma_0^2\right)^2} & -\frac{E(x_{1t}x_{2t})\gamma_0}{2\left(\sigma_0^2\right)^2} + \frac{3E[x_{1t}x_{2t}\gamma_0] + E\left[x_{1t}(x_{2t}\gamma_0)^3\right]}{2\left(\sigma_0^2\right)^2} \\ -\frac{\gamma_0 E(x_{1t}x_{2t})}{2\left(\sigma_0^2\right)^2} + \frac{3\sigma_0^2 E[\gamma_0 x_{1t}x_{2t}] + E\left[(x_{2t}\gamma_0)^3 x_{1t}\right]}{2\left(\sigma_0^2\right)^3} & \frac{1}{2\left(\sigma_0^2\right)^2} - \frac{E(x_{2t}\gamma_0)^2}{2\left(\sigma_0^2\right)^3} + \frac{6\sigma_0^2 E(x_{2t}\gamma_0)^2 + E(x_{2t}\gamma_0)^4}{4\left(\sigma_0^2\right)^4} \end{bmatrix} \\ & -\frac{1}{n}\sum_{t=1}^{n}\begin{bmatrix} \frac{[\gamma_0 E(x_{1t}x_{2t})]^2}{\left(\sigma_0^2\right)^2} & \frac{E(x_{1t}x_{2t})\gamma_0 E\left[(x_{2t}\gamma_0)^2\right]}{2\left(\sigma_0^2\right)^3} \\ \frac{E\left[(x_{2t}\gamma_0)^2\right]\gamma_0 E(x_{1t}x_{2t})}{2\left(\sigma_0^2\right)^3} & \frac{\left(E\left[(x_{2t}\gamma_0)^2\right]\right)^2}{4\left(\sigma_0^2\right)^4} \end{bmatrix}. \end{aligned}$$

Since $\gamma_0 = \delta_0/n^{1/2}$, *we have*

$$\lim_{n\to\infty}\mathbf{B}_n\left(\boldsymbol{\theta}^t\right) = \lim_{n\to\infty} -\mathbf{H}_n\left(\boldsymbol{\theta}^t\right) = \begin{bmatrix} \lim_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\frac{E\left(x_{1t}^2\right)}{\sigma_0^2} & 0 \\ 0 & \frac{1}{2\left(\sigma_0^2\right)^2} \end{bmatrix}.$$

This means that $\mathbf{H}_n\left(\boldsymbol{\theta}^t\right) = -\mathbf{B}_n\left(\boldsymbol{\theta}^t\right) + o\left(1\right)$. *However, we do not have* $\mathbf{H}_n\left(\boldsymbol{\theta}^t\right) = -\mathbf{B}_n\left(\boldsymbol{\theta}^t\right)$ *for any finite* n.

Lemma 3.1 *Under Assumptions 1-15, we have*

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

Remark 3.4 *Under different regularity conditions, the Bernstein-von Mises theorem shows that the posterior distribution converges to a normal distribution with the QML estimator as its mean and the inverse of the second derivative of the log-likelihood function evaluated at the QML estimator as its covariance. Based on Bernstein-von Mises theorem, Ghosh and Ramamoorthi (2003) developed the same results as in Lemma 3.1 for the iid case. We extend the results of Ghosh and Ramamoorthi (2003) to more general cases.*

Theorem 3.1 *Under Assumptions 1-15, when the prior of $\boldsymbol{\theta}$ is $O_p(1)$, we have,*

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

Remark 3.5 *Like AIC, DIC is an unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ asymptotically, according to Theorem 3.1. Hence, the decision-theoretic justification to DIC is that DIC selects a model that minimizes the frequentist risk, which is the expected KL divergence between the DGP and the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$ where the expectation is taken with respect to the DGP. A key difference between AIC and DIC is that the plug-in predictive distribution is based on different estimators. In AIC the QML estimate, $\hat{\boldsymbol{\theta}}(\mathbf{y})$, is used while in DIC the Bayesian posterior mean, $\bar{\boldsymbol{\theta}}(\mathbf{y})$, is used.*

Remark 3.6 *The justification of DIC remains valid if the posterior mean replaced with the posterior mode or with the QML estimate and P_D is replaced with P. This is because the justification of DIC requires that the information matrix identity holds true asymptotically, and that the posterior distribution converges to a normal distribution (the posterior mean converges to the posterior mode and the posterior variance converges to zero). That is why DIC is explained as the Bayesian version of AIC.*

Remark 3.7 *In AIC, the number of degrees of freedom, P, is used to measure the model complexity. In the Bayesian framework, the prior information often imposes additional restrictions on the parameter space and, hence, the degrees of freedom may be reduced by the usage of a prior. In this case, P_D may not be close to P for a finite n. A useful contribution of DIC is to provide a way to measure the model complexity when the prior information is incorporated; see Brooks (2002).*

Remark 3.8 *As pointed out in Spiegelhalter, et al (2014), the consistency of BF requires that there is a "true model" and the "true model" is among the candidate models. However, AIC and DIC are prediction-based criteria which are designed to find the best model for making predictions among candidate models. Neither AIC nor DIC makes attempt to find the "true model".*

Remark 3.9 *If $p(\mathbf{y}|\boldsymbol{\theta})$ has a closed-form expression, DIC is trivially computable from the MCMC output. The computational tractability, together with the versatility of MCMC and the fact that DIC is incorporated into a Bayesian software, WinBUGS, allows DIC to enjoy a very wide range of applications.*

4 DIC Based on Bayesian Predictive Distribution

The above decision-theoretic justification of DIC is based on the loss function constructed from the plug-in predictive distribution. Unfortunately, the plug-in predictive distribution

is not invariant to parameterization and, hence, the corresponding DIC can be sensitive to parameterization. From the pure Bayesian viewpoint, only the Bayesian predictive distribution, but not the plug-in predictive distribution, is a full proper predictive distribution. The Bayesian predictive distribution is invariant to reparameterization. Hence, the loss function and the corresponding information criterion will be invariant to reparameterization; see Ando and Tsai (2010), Spiegelhalter, et al (2014).

Let $p(\mathbf{y}_{rep}|\mathbf{y},M_k)$ be the Bayesian predictive distribution, that is,

$$p(\mathbf{y}_{rep}|\mathbf{y},M_k) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta},M_k)p(\boldsymbol{\theta}|\mathbf{y},M_k)d\boldsymbol{\theta}.$$

The KL divergence based on the Bayesian predictive distribution is

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

The frequentist risk for a statistical decision d_k that selects Model M_k is

$$\begin{aligned} Risk(d_k) &= E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right),p\left(\mathbf{y}_{rep}|\mathbf{y},M_k\right)\right]\right\} \\ &= E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left(2\ln g\left(\mathbf{y}_{rep}\right)\right)\right\} + E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left(-2p\left(\mathbf{y}_{rep}|\mathbf{y},M_k\right)\right)\right\}. \end{aligned}$$

A better model is expected to yield a smaller value for the $Risk(d_k)$. Since $E_{\mathbf{y}_{rep}}\left(2\ln g\left(\mathbf{y}_{rep}\right)\right)$ is the same across all candidate models, it is dropped from (10) when comparing models. As a result, we propose to choose a model that gives the smallest value of (again we suppress M_k for notational simplicity)

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

Let the selected model be M_{k^*} (i.e. the optimal decision is d_{k^*}). Then $p(\mathbf{y}_{rep}|\mathbf{y},M_{k^*})$, which is the closest to $g\left(\mathbf{y}_{rep}\right)$ in terms of the expected KL divergence, is used to generate predictions of future observations.

4.1 $\mathbf{IC}_{AT}$

Under the iid assumption, Ando and Tsai (2010) showed that

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] = E_{\mathbf{y}}\left\{-2\ln p(\mathbf{y}|\mathbf{y}) + \mathbf{tr}\left[\mathbf{J}_{AT}^{-1}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\mathbf{I}_{AT}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\right]\right\} + o(1),$$

where

$$\hat{\boldsymbol{\theta}}_{AT} = \arg\max_{\boldsymbol{\theta}\in\Theta} 2\ln p(\mathbf{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta}),$$

$$\mathbf{J}_{AT}(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^{n}\left\{\frac{\partial^2\ln\xi(y_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right\}, \mathbf{I}_{AT}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n}\left\{\frac{\partial\ln\xi(y_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\frac{\partial\ln\xi(y_t|\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\right\},$$

with $\ln\xi(y_t|\boldsymbol{\theta}) = \ln p(y_t|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})/(2n)$ for $t = 1,\cdots,n$.

Remark 4.1 *Ando and Tsai (2010) defined the following information criterion*

$$IC_{AT} = -2\ln p(\mathbf{y}|\mathbf{y}) + \mathbf{tr}\left[\mathbf{I}_{AT}^{-1}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\mathbf{J}_{AT}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\right]. \tag{11}$$

Since IC_{AT} is constructed based on the Bayesian predictive distribution, it is invariant to reparameterization. Interestingly, the first term in IC_{AT} is different from that in AIC or DIC. To compute the first term in IC_{AT} from the MCMC output, note that

$$-2\ln p(\mathbf{y}|\mathbf{y}) = -2\ln\left(\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\right) \approx -2\ln\left(\frac{1}{J}\sum_{j=1}^{J} p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)\right),$$

where $\left\{\boldsymbol{\theta}^{(j)}\right\}_{j=1}^{J}$ are J effective random samples drawn from the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. To compute the penalty term $\mathbf{tr}\left[\mathbf{I}_{AT}^{-1}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\mathbf{J}_{AT}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\right]$, one first needs to obtain $\hat{\boldsymbol{\theta}}_{AT}$. In general, $\hat{\boldsymbol{\theta}}_{AT}$ is not available analytically and hence, one has to use a numerical optimizer to find $\hat{\boldsymbol{\theta}}_{AT}$. Then one needs to calculate $\mathbf{I}_{AT}(\boldsymbol{\theta})$ and $\mathbf{J}_{AT}(\boldsymbol{\theta})$ and to invert $\mathbf{I}_{AT}(\boldsymbol{\theta})$.

Remark 4.2 *Under the assumptions that the prior is $O_p(1)$ and that the candidate model encompasses the DGP, Ando and Tsai simplified the information criterion as*

$$IC_{AT} = -2\ln p(\mathbf{y}|\mathbf{y}) + P. \tag{12}$$

In this case, the second term has a very simple expression as it is the same as the number of degrees of freedom, which no longer depends on the prior information.

4.2 DIC^{BP}

Using $-2\ln p(\mathbf{y}|\mathbf{y})$ as the first term makes it difficult to compare with DIC, AIC or BIC. In this paper, we propose a Bayesian predictive distribution-based information criterion whose first term is the same as DIC, i.e., $D\left(\bar{\boldsymbol{\theta}}\right)$. It turns out such a choice leads to a simple penalty term and facilitates comparison with DIC, AIC or BIC. In the following theorem we propose a new information criterion based on the KL divergence between the DGP and the Bayesian predictive distribution and show the asymptotic unbiasedness.

Theorem 4.1 *Define the information criterion based on the Bayesian predictive distribution as*

$$DIC^{BP} = D\left(\bar{\boldsymbol{\theta}}\right) + (1+\ln 2)P_D, \tag{13}$$

where P_D is defined in (4). Under Assumptions 1-15 and when the prior of $\boldsymbol{\theta}$ is $O_p(1)$, we have

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

Remark 4.3 *The justification of DIC^{BP} remains valid if the posterior mean is replaced with the posterior mode or with the QML estimate and if P_D is replaced with P. Clearly, the penalty term in DIC^{BP} is smaller than that in DIC, AIC, and BIC (i.e., $(1+\ln 2)P_D \approx (1+\ln 2)P$ as opposed to $2P_D$ in DIC, $2P$ in AIC, and $P\ln n$ in BIC).*

Remark 4.4 *It can be shown that $DIC^{BP} = IC_{AT} + o_p(1)$ and $E_{\mathbf{y}}\left[DIC^{BP}\right] = E_{\mathbf{y}}(IC_{AT}) + o(1)$. Clearly $(1+\ln 2)P_D \approx (1+\ln 2)P > P$, implying $-2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) < -2\ln p(\mathbf{y}|\mathbf{y})$. To see why this is the case, note that $p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) = p\left(\mathbf{y}|\left(\int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\right)\right)$, and $p(\mathbf{y}|\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}$. Under Assumptions 2-15, $p(\boldsymbol{\theta}|\mathbf{y})$ is approximately Gaussian and concave. The inequality $-2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) < -2\ln p(\mathbf{y}|\mathbf{y})$ follows from Jensen's inequality.*

Remark 4.5 *As IC_{AT}, DIC^{BP} is based on the Bayesian predictive distribution and hence, is invariant to reparameterization. There are several good properties for DIC^{BP}. First, DIC^{BP} is developed without resorting the iid assumption. Second, DIC^{BP} is easier to compute. When the MCMC output is available, IC_{AT} needs to evaluate*

$$\ln p(\mathbf{y}|\mathbf{y}) = \ln\left(\int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\right) \approx \ln\left(\frac{1}{J}\sum_{j=1}^{J} p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)\right),$$

while DIC^{BP} needs to evaluate

$$\int \ln p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta} \approx \frac{1}{J}\sum_{j=1}^{J} \ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right).$$

Numerically, the log-likelihood function is usually much more stable than the likelihood function. Thus, for the same value of J, the accuracy in $\frac{1}{J}\sum_{j=1}^{J}\ln p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)$ is often much higher than that in $\ln\left(\frac{1}{J}\sum_{j=1}^{J} p\left(\mathbf{y}|\boldsymbol{\theta}^{(j)}\right)\right)$. Moreover, DIC^{BP} is as easy to compute as DIC. Since DIC is monitored in WinBUGS, no additional effort is needed for calculating DIC^{BP}. Third, like DIC, the penalty term depends on the prior information. As pointed out by Brooks (2002), a useful contribution of DIC is to provide a way to measure the model complexity when the prior information is incorporated. This property is shared by DIC^{BP}.

Remark 4.6 *A recent literature suggests the minimization of the posterior mean of the KL divergence between $g(\mathbf{y}_{rep})$ $p(\mathbf{y}_{rep}|\boldsymbol{\theta})$, i.e.,*

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

Hence, the corresponding frequentist risk for a statistical decision d_k is

$$\begin{aligned} & Risk(d_k) = E_{\mathbf{y}}\left\{\int\left[\int \ln\frac{g(\mathbf{y}_{rep})}{p(\mathbf{y}_{rep}|\boldsymbol{\theta},M_k)} g(\mathbf{y}_{rep})\,d\mathbf{y}_{rep}\right] p(\boldsymbol{\theta}|\mathbf{y},M_k)\, d\boldsymbol{\theta}\right\} \\ = & \int \ln g(\mathbf{y}_{rep})\, g(\mathbf{y}_{rep})\, d\mathbf{y}_{rep} - E_{\mathbf{y}}\left\{\int\int \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta},M_k)\, g(\mathbf{y}_{rep})\, d\mathbf{y}_{rep} p(\boldsymbol{\theta}|\mathbf{y},M_k)\, d\boldsymbol{\theta}\right\}. \end{aligned}$$

Since the first term is constant across different models, van der Linde (2005, 2012), Plummer (2008), Ando (2007) and Ando (2012) proposed to choose a model to minimize

$$E_{\mathbf{y}}\left\{\int\int\left[-2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)g\left(\mathbf{y}_{rep}\right)d\mathbf{y}_{rep}\right]p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}\right\}.$$

Under the iid assumption, it was shown that

$$E_{\mathbf{y}}\left[\int\int\left[-2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)g\left(\mathbf{y}_{rep}\right)d\mathbf{y}_{rep}\right]p(\boldsymbol{\theta}|\mathbf{y})d\boldsymbol{\theta}\right]\approx E_{\mathbf{y}}\left[D\left(\bar{\boldsymbol{\theta}}\right)+3P_D\right],$$

leading to an information criterion called BPIC by Ando (2007). Clearly, the target is different here from that under DIC and also from that under DIC^{BP}. According to Spiegelhalter, et al (2014), BPIC chooses an "average target" rather than a "representative target". More importantly, although BPIC can select a model, it cannot tell the user how to actually predict the future observations.

Remark 4.7 *To understand why the penalty term in BPIC is larger than that in DIC^{BP}, note that the loss employed by BPIC is*

$$\begin{aligned}&\int\int-2\ln\left(p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right)g\left(\mathbf{y}_{rep}\right)p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\,d\mathbf{y}_{rep}\\ =\;&\int\int-2\ln\left(p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right)p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}g\left(\mathbf{y}_{rep}\right)d\mathbf{y}_{rep},\end{aligned}$$

while the loss by DIC^{BP} is

$$\int-2\ln\left(\int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\right)g\left(\mathbf{y}_{rep}\right)d\mathbf{y}_{rep}.$$

By Jensen's inequality,

$$-2\ln\left(\int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}\right)<\int-2\ln\left(p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)\right)p(\boldsymbol{\theta}|\mathbf{y})\,d\boldsymbol{\theta}.$$

Hence, the frequentist risk in DIC^{BP} is smaller than that in BPIC for the same candidate model.

4.3 Frequentist Risk of DIC and DIC^{BP}

From the decision viewpoint, DIC and DIC^{BP} lead to different statistical decisions. If the optimal model selected by DIC is different from that selected by DIC^{BP}, the two statistical decisions are obviously different. Even when DIC and DIC^{BP} select the same model , the two statistical decisions are still different because the predictions come from two different distributions, namely $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y}),M_k\right)$ versus $p\left(\mathbf{y}_{rep}|\mathbf{y},M_k\right)$. In both cases, therefore, the frequentist risk implied by DIC and DIC^{BP} is different. Although it is known that the Bayesian predictive distribution is a full predictive distribution and invariant to parameterization, the

questions such as "which predictive distribution should be used for making predictions?" and "is there any difference in using the two predictive distributions for making predictions?" remain unanswered in the literature. The theoretical development of DIC^{BP} allows us to answer these two important questions.

With the two information criteria, the action space is larger than before. Denote the action space by $\{d_{k^0}, d_{k^1}\}_{k=1}^{K}$ where d_{k^a} $(a \in (0,1))$ means M_k is selected, and the predictions come from $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)$ if $a = 0$ (i.e., DIC is the corresponding information criterion) but the predictions come from $p\left(\mathbf{y}_{rep}|\mathbf{y}, M_k\right)$ if $a = 1$ (i.e., DIC^{BP} is the corresponding information criterion). Let the two KL divergence functions be represented uniformly as

$$\mathcal{L}(\mathbf{y}, d_{k^a}) = 2 \times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\mathbf{y}, d_{k^a}\right)\right],$$

where $p\left(\mathbf{y}_{rep}|\mathbf{y}, d_{k^0}\right) := p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)$, and $p\left(\mathbf{y}_{rep}|\mathbf{y}, d_{k^1}\right) := p\left(\mathbf{y}_{rep}|\mathbf{y}, M_k\right)$. The frequentist risk associated with d_{k^a} is

$$Risk(d_{k^a}) = E_{\mathbf{y}}\left(\mathcal{L}(\mathbf{y}, d_{k^a})\right) = \int \mathcal{L}(\mathbf{y}, d_{k^a}) g(\mathbf{y}) \mathrm{d}\mathbf{y}.$$

Hence, the model selection problem is equivalent to the following statistical decision,

$$\min_{a\in\{0,1\}} \min_{k\in\{1,\cdots,K\}} Risk\left(d_{k^a}\right). \tag{14}$$

According to Equation (19) in Appendix, $P_D = P + o_p(1)$. Since $P > 0$ and $\ln 2 + 1 < 2$, for any model M_k, we have $\text{DIC} > \text{DIC}^{BP}$ with probability approaching one (w.p.a.1). As a result, $E_{\mathbf{y}}\left(\text{DIC}\right) > E_{\mathbf{y}}\left(\text{DIC}^{BP}\right)$ w.p.a.1. Following Theorem 3.1 and Theorem 4.1, w.p.a.1, we have

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

$$2E_{\mathbf{y}} E_{\mathbf{y}_{rep}} g(\mathbf{y}_{rep}) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_k(\mathbf{y}), M_k\right)\right] > 2E_{\mathbf{y}} E_{\mathbf{y}_{rep}} g(\mathbf{y}_{rep}) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}, M_k\right)\right],$$

and

$$Risk\left(d_{k^0}\right) = E_{\mathbf{y}}\left(\mathcal{L}\left(\mathbf{y}, d_{k^0}\right)\right) > E_{\mathbf{y}}\left(\mathcal{L}\left(\mathbf{y}, d_{k^1}\right)\right) = Risk\left(d_{k^1}\right). \tag{15}$$

Hence, w.p.a.1,

$$\min_{k\in\{1,\cdots,K\}} Risk(d_{k^0}) > \min_{k\in\{1,\cdots,K\}} Risk\left(d_{k^1}\right). \tag{16}$$

and

$$\arg\min_{a\in\{0,1\}}\left[\min_{k\in\{1,\cdots,K\}} Risk(d_{k^a})\right] = 1.$$

This means that the optimal solution to the statistical decision problem given in Section 2.1 is obtained by DIC^{BP} w.p.a.1. This is true even in case where $\arg\min_{k\in\{1,\cdots,K\}} Risk(d_{k^0}) = \arg\min_{k\in\{1,\cdots,K\}} Risk(d_{k^1})$. Therefore, as far as the frequentist risk is concerned, the Bayesian predictive distribution but not the plug-in predictive distribution should be used for making predictions.

5 Conclusion

This paper provides a rigorous decision-theoretic justification of DIC based on a set of regularity conditions but without requiring the iid assumption. The candidate model is not required to encompassed the DGP. It is shown that DIC is an asymptotically unbiased estimator of the expected KL divergence between the DGP and the plug-in predictive distribution.

Based on the Bayesian predictive distribution, a new information criterion (DIC^{BP}) is constructed for model selection. The first term has the same expression as that in DIC, but the penalty term is smaller than that in DIC. The asymptotic justification of DIC^{BP} is provided, in the same way as how DIC has been justified. The frequentist risk of DIC^{BP} is compared with that of DIC and BPIC. It is shown that as $n \to \infty$, DIC^{BP} leads to the smaller frequentist risk than DIC and BPIC. From the decision viewpoint, the Bayesian predictive distribution and DIC^{BP} but not the plug-in predictive distribution or DIC leads to the optimal decision action.

Although the theoretic framework under which we justify DIC and DIC^{BP} are general, it requires the consistency of the posterior mean, the asymptotic normal approximation to the posterior distribution, and the asymptotic normality to the QML estimator. When there are latent variables in the candidate model under which the number of latent variables grows as n grows and when the parameter space is enlarged to include latent variables, the consistency and the asymptotic normality may not hold true. As a result, DIC and DIC^{BP} are not justified. Moreover, when the data are nonstationary, the asymptotic normality may not hold true. In this case, it remains unknown whether or not DIC and DIC^{BP} are still justified.

Appendix

Notations

$:=$	definitional Equality	$\overleftarrow{\boldsymbol{\theta}}$	posterior mode
$o(1)$	tend to zero	$\widehat{\boldsymbol{\theta}}$	QML estimate
$o_p(1)$	tend to zero in probability	$\boldsymbol{\theta}^t$	pseudo true parameter
$\xrightarrow{p}$	converge in probability	$\hat{\boldsymbol{\theta}}_{AT}$	$\arg\max$ of $2\ln p(\mathbf{y}\|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$
$\overline{\boldsymbol{\theta}}$	posterior mean	$\widetilde{\boldsymbol{\theta}}$	$\arg\max$ of $\ln p\left(\mathbf{y}_{rep}\|\boldsymbol{\theta}\right) + \ln p\left(\mathbf{y}\|\boldsymbol{\theta}\right) + \ln p\left(\boldsymbol{\theta}\right)$

Proof of Lemma 3.1

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

$$\begin{aligned}
\ln p(\boldsymbol{\theta}|\mathbf{y}) &= \ln p\left(\overleftarrow{\boldsymbol{\theta}}|\mathbf{y}\right) + L_n^{(1)}\left(\overleftarrow{\boldsymbol{\theta}}\right)'\left(\boldsymbol{\theta} - \overleftarrow{\boldsymbol{\theta}}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \overleftarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\tilde{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta} - \overleftarrow{\boldsymbol{\theta}}\right) \\
&= \ln p\left(\overleftarrow{\boldsymbol{\theta}}|\mathbf{y}\right) + \frac{1}{2}\left(\boldsymbol{\theta} - \overleftarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\tilde{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta} - \overleftarrow{\boldsymbol{\theta}}\right),
\end{aligned}$$

where $\tilde{\boldsymbol{\theta}}_1$ lies on the segment between $\boldsymbol{\theta}$ and $\overleftrightarrow{\boldsymbol{\theta}}$. The Taylor expansion for a random function is justified by Lemma 3 of Jennrich (1969). It follows that

$$p(\boldsymbol{\theta}|\mathbf{y}) = p\left(\overleftrightarrow{\boldsymbol{\theta}}|\mathbf{y}\right) \exp\left[\frac{1}{2}\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\tilde{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)\right].$$

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

$$p(\boldsymbol{\omega}|\mathbf{y}) \propto \exp\left[\frac{1}{2}\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\tilde{\boldsymbol{\theta}}_1\right)\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)\right] = \exp\left\{-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right\}.$$

Letting $c_n^* = \int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}$, $c_n = \int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}$, we have

$$\begin{aligned}
& P_n := \int_\Omega \left| p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \int_\Omega \left| \frac{1}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] - \frac{1}{c_n}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \frac{1}{c_n}\int_\Omega \left| \frac{c_n}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
= & \frac{1}{c_n}\int_\Omega \left| \frac{c_n - c_n^*}{c_n^*}\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] + \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{1}{c_n}\left\{\int_\Omega \left|\frac{c_n - c_n^*}{c_n^*}\right| \exp\left[-\frac{\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}}{2}\right] \mathrm{d}\boldsymbol{\omega} + \int_\Omega \left|\exp\left[\frac{\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}}{2}\right] - \exp\left[-\frac{\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2}\right]\right| \mathrm{d}\boldsymbol{\omega}\right\} \\
\leq & \frac{|c_n - c_n^*|}{c_n} + \frac{1}{c_n}\int_\Omega \left|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{2}{c_n}\int_\Omega \left|\exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] - \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right]\right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \frac{2}{c_n}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right]\boldsymbol{\omega}\right\} - 1\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

When $\Omega = \{\boldsymbol{\omega} : ||\boldsymbol{\omega}|| < \sqrt{n}\delta\}$, we have $\boldsymbol{\theta} \in H(\overleftrightarrow{\boldsymbol{\theta}}, \delta)$ and $-A(\epsilon) \leq [J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P] \leq A(\epsilon)$. Define

$$Q_n = \int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right]\boldsymbol{\omega}\right\} - 1\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}.$$

By the Hölder inequality, we have

$$\begin{aligned}
&\lim_{n\to\infty} Q_n = \lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right]\boldsymbol{\omega}\right\} - 1\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega} \\
= &\lim_{n\to\infty}\int_\Omega \left|\exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P\right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right\} - 1\right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega} \\
\leq &\lim_{n\to\infty} c_n \left\{\int_\Omega \left|\exp\left\{-\frac{\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P\right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2}\right\} - 1\right|^2 \frac{1}{c_n}\exp\left[-\frac{\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2}\right] \mathrm{d}\boldsymbol{\omega}\right\}^{1/2} \\
= &\ D_1^{1/2}(D_1 - 2D_2 + D_3)^{1/2},
\end{aligned}$$

where

$$\begin{aligned}
D_1 &= \lim_{n\to\infty}\int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}, \\
D_2 &= \lim_{n\to\infty}\int_\Omega \exp\left\{-\frac{1}{2}\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P\right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right\} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega} \\
&= \lim_{n\to\infty}\int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}, \\
D_3 &= \lim_{n\to\infty}\int_\Omega \exp\left\{-\boldsymbol{\omega}'\left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P\right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right\} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}\right] \mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

It can be shown that $D_1 = \lim_{n\to\infty} c_n = (2\pi)^{P/2}|J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)|^{-1/2}$. Following the proof of Lemma 2.1 and Theorem 2.1 of Chen (1985), we have $D_2^- \leq D_2 \leq D_2^+, D_3^- \leq D_3 \leq D_3^+$ and

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

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

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

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

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

$$\begin{aligned}
& \lim_{n\to\infty} \left| \int_\Omega \omega_i \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \right\} \mathrm{d}\boldsymbol{\omega} \right| \\
\leq & \lim_{n\to\infty} \int_\Omega \left| \omega_i \left\{ p(\boldsymbol{\omega}|\mathbf{y}) - \frac{1}{c_n} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \right\} \right| \mathrm{d}\boldsymbol{\omega} \\
\leq & \lim_{n\to\infty} \frac{|c_n - c_n^*|}{c_n} \int |\omega_i| p(\boldsymbol{\omega}|\mathbf{y}) \mathrm{d}\boldsymbol{\omega} \\
+ & \lim_{n\to\infty} \frac{1}{c_n} \int_\Omega |\omega_i| \left| \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \right] \boldsymbol{\omega} \right\} - 1 \right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}.
\end{aligned}$$

By Assumption 5, we have

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

By Hölder's inequality, we have

$$\begin{aligned}
& \int_\Omega |\omega_i| \left| \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \right] \boldsymbol{\omega} \right\} - 1 \right| \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
= & c_n \int_\Omega |\omega_i| \left| \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \right] \boldsymbol{\omega} \right\} - 1 \right| \frac{1}{c_n} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
\leq & \left\{ \int_\Omega \left| \exp\left\{ -\frac{\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P \right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2} \right\} - 1 \right|^2 \exp\left[-\frac{\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2} \right] \mathrm{d}\boldsymbol{\omega} \right\}^{\frac{1}{2}} \\
& \times \left\{ \int_\Omega \omega_i^2 \exp\left[-\frac{\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega}}{2} \right] \mathrm{d}\boldsymbol{\omega} \right\}^{\frac{1}{2}} \\
= & \sqrt{E\omega_i^2} (ED_1 - 2ED_2 + ED_3)^{1/2} \longrightarrow 0,
\end{aligned}$$

where

$$\begin{aligned}
E(D_1) &= \int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}, \\
E(D_2) &= \int_\Omega \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P \right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right\} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
&= \int_\Omega \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\tilde{\boldsymbol{\theta}}_1\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega}, \\
E(D_3) &= \int_\Omega \omega_i^2 \exp\left\{ -\boldsymbol{\omega}' \left[J\left(\tilde{\boldsymbol{\theta}}_1\right) J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) - I_P \right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right\} \exp\left[-\frac{1}{2}\boldsymbol{\omega}' J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right] \mathrm{d}\boldsymbol{\omega} \\
= & \int_\Omega \omega_i^2 \exp\left\{ -\frac{1}{2}\boldsymbol{\omega}' \left[2J\left(\tilde{\boldsymbol{\theta}}_1\right) - J\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \right] J\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\boldsymbol{\omega} \right\} \mathrm{d}\boldsymbol{\omega}
\end{aligned}$$

and $E(D_1) - 2E(D_2) + E(D_3) \longrightarrow 0$

Hence, we have

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

Similarly, we also can show that

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

Note that

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

where $J_{ij}^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}} \right)$ is the $(i,j)^{th}$ element of $J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}} \right)$. Hence, $E(\boldsymbol{\omega}|\mathbf{y}) = 0 + o_p(1)$ and $E\left(\boldsymbol{\omega}\boldsymbol{\omega}' | \mathbf{y} \right) = J^{-1}\left(\overleftrightarrow{\boldsymbol{\theta}} \right) + o_p(1)$ which imply that

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

Proof of Theorem 3.1

Following Bester and Hansen (2006), under Assumption 1-15, we have

$$\overleftrightarrow{\boldsymbol{\theta}}(\mathbf{y}) = \overline{\boldsymbol{\theta}}(\mathbf{y}) + O_p\left(n^{-1} \right).$$

Then we can show that

$$\overline{\boldsymbol{\theta}}(\mathbf{y}) = \boldsymbol{\theta}^t + O_p\left(n^{-1/2} \right),$$

$$\frac{1}{\sqrt{n}} B_n^{-1} \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}^t)}{\partial \boldsymbol{\theta}} \xrightarrow{d} N(0, I), \tag{17}$$

and

$$C_n^{-1/2} \sqrt{n} \left(\overline{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^t \right) \xrightarrow{d} \mathcal{N}(0, I_P), \tag{18}$$

where $C_n = H_n B_n H_n^{-1}$. When there is no confusion, we write $H_n\left(\boldsymbol{\theta}^t \right)$ as H_n and $B_n\left(\boldsymbol{\theta}^t \right)$ as B_n.

Note that

$$\begin{aligned}
& E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}(\mathbf{y})\right)\right) \\
= & \underset{(T_1)}{\left[E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)\right)\right]} \\
+ & \underset{(T_2)}{\left[E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t\right)\right) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)\right)\right]} \\
+ & \underset{(T_3)}{\left[E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}(\mathbf{y})\right)\right) - E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left(-2 \ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t\right)\right)\right]}.
\end{aligned}$$

Now let us analyze T_2 and T_3. First expand $\ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t\right)$ at $\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)$,

$$\begin{aligned}
\ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t\right) &= \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right) + \frac{\partial \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)'}{\partial \boldsymbol{\theta}} \left(\boldsymbol{\theta}^t - \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right) \\
&\quad + \left(\boldsymbol{\theta}^t - \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)' \frac{\partial \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\boldsymbol{\theta}^t - \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right) + o_p(1) \\
&= \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right) + \left(\boldsymbol{\theta}^t - \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)' \frac{\partial \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\boldsymbol{\theta}^t - \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right) + o_p(1).
\end{aligned}$$

Then we have

$$\begin{aligned}
T_2 &= E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[-2 \ln p \left(\mathbf{y}_{rep} | \boldsymbol{\theta}^t\right) + 2 \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)\right] \\
&= E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \left[- \left(\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}^t\right)' \frac{\partial \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}^t\right) + o_p(1)\right] \\
&= E_{\mathbf{y}_{rep}} \left[- \left(\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}^t\right)' \frac{\partial \ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}^t\right)\right] + o(1) \\
&= E_{\mathbf{y}} \left[- \left(\overline{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^t\right)' \frac{\partial \ln p(\mathbf{y} | \overline{\boldsymbol{\theta}}(\mathbf{y}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \left(\overline{\boldsymbol{\theta}}(\mathbf{y}) - \boldsymbol{\theta}^t\right)\right] + o(1),
\end{aligned}$$

by Assumption 10 and the dominated convergence theorem. Next, we expand $\ln p \left(\mathbf{y}_{rep} | \overline{\boldsymbol{\theta}}(\mathbf{y})\right)$ at $\boldsymbol{\theta}^t$:

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

Substituting the above expansion into T_3, we have

$$\begin{aligned}
T_3 &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}(\mathbf{y})\right)) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)) \\
&= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}}'\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right) - \left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right) + o_p(1)\right] \\
&= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}}'\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] \\
&\quad + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] + o(1) \\
&= -2E_{\mathbf{y}_{rep}}\left(\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}}\right)' E_{\mathbf{y}}\left[\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] \\
&\quad + E_{\mathbf{y}}\left[-\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'E_{\mathbf{y}_{rep}}\left(\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] + o(1) \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'E_{\mathbf{y}}\left(\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}^t)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] + o(1),
\end{aligned}$$

since

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

by (17), (18), and the dominated convergence theorem.

Note that

$$\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}(\mathbf{y})\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'} = E_{\mathbf{y}}\left(\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right) + o_p(1),$$

under Assumption 6-15 and by the uniform law of large number. Hence, we get

$$\begin{aligned}
T_2 &= E_{\mathbf{y}}\left[-\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\frac{\partial \ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}(\mathbf{y})\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right] + o(1) \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'\frac{1}{n}E_{\mathbf{y}}\left(\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right) + o_p(1)\right] + o(1) \\
&= T_3 + o(1).
\end{aligned}$$

So we only need to analyze T_3. Note that

$$\begin{aligned}
T_3 &= E_{\mathbf{y}}\left[-\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)' E_{\mathbf{y}}\left(-\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}^t)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]+o(1)\\
&= E_{\mathbf{y}}\left[\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)'(-H_n)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]+o(1)\\
&= E_{\mathbf{y}}\left[\left(C_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right)' C_n^{1/2}(-H_n)C_n^{1/2}C_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\right]+o(1)\\
&= E_{\mathbf{y}}\left\{\mathbf{tr}\left[H_nC_n^{1/2}C_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)' C_n^{-1/2}C_n^{1/2}\right]\right\}+o(1)\\
&= \mathbf{tr}\left\{(-H_n)C_n^{1/2}E_{\mathbf{y}}\left[C_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)' C_n^{-1/2}\right]C_n^{1/2}\right\}+o(1)\\
&= \mathbf{tr}\left\{(-H_n)C_n^{1/2}E_{\mathbf{y}}\left[C_n^{-1/2}\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\sqrt{n}\left(\widehat{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)' C_n^{-1/2}\right]C_n^{1/2}\right\}+o(1),
\end{aligned}$$

Then, we have

$$E_{\mathbf{y}}\left[C_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}(\mathbf{y})-\boldsymbol{\theta}^t\right)' C_n^{-1/2}\right]=I_P+o(1).$$

Hence,

$$\begin{aligned}
T_3 &= \mathbf{tr}\left((-H_n)C_n^{1/2}C_n^{1/2}\right)+o(1)=\mathbf{tr}\left((-H_n)C_n\right)+o(1)\\
&= \mathbf{tr}\left((-H_n)(-H_n)^{-1}B_n(-H_n)^{-1}\right)+o(1)=\mathbf{tr}\left((-H_n)(-H_n)^{-1}B_n(-H_n)^{-1}\right)+o(1)\\
&= \mathbf{tr}\left(B_n(-H_n)^{-1}\right)+o(1),
\end{aligned}$$

and

$$\begin{aligned}
E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}(-2\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}(\mathbf{y})\right))\right] &= E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}(-2\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)\right))\right]+2\mathbf{tr}\left(B_n(-H_n)^{-1}\right)+o(1)\\
&= E_{\mathbf{y}}\left[E_{\mathbf{y}}(-2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}(\mathbf{y})))\right]+2\mathbf{tr}\left(B_n(-H_n)^{-1}\right)+o(1)\\
&= E_{\mathbf{y}}(-2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}(\mathbf{y})))+2\mathbf{tr}\left(B_n(-H_n)^{-1}\right)+o(1)\\
&= E_{\mathbf{y}}(-2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}(\mathbf{y}))+2P)+o(1),
\end{aligned}$$

under the condition that $B_n=-H_n+o(1)$. If the candidate model is correctly specified, $B_n=-H_n$ for each n, then the condition is automatically satisfied.

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

$$\begin{aligned}
&\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right)=\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)+\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)'}{\partial\theta}\left(\overline{\boldsymbol{\theta}}-\overleftrightarrow{\boldsymbol{\theta}}\right)+\frac{1}{2}\left(\overline{\boldsymbol{\theta}}-\overleftrightarrow{\boldsymbol{\theta}}\right)'\frac{\partial^2 \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_2\right)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}-\overleftrightarrow{\boldsymbol{\theta}}\right)\\
= &\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)+O_p(n^{1/2})O_p(n^{-1})+O_p(n^{-1})O_p(n)o_p(n^{-1})\\
= &\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)+O_p(n^{-1/2}),
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_2$ lies on the segment between $\overline{\boldsymbol{\theta}}$ and $\overleftrightarrow{\boldsymbol{\theta}}$.

Following Lemma 3.1, we get

$$\begin{aligned}
& P_D = \int -2\left[\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right)\right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
= & \int -2\left[\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} + 2\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) - 2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) \\
= & -2\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)'}{\partial\theta}\left(\bar{\boldsymbol{\theta}} - \overleftrightarrow{\boldsymbol{\theta}}\right) - \int \left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' \frac{\partial^2 \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_3\right)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right) p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} + o_p(1) \\
= & o_p(1) - \int \left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\widetilde{\boldsymbol{\theta}}_3\right)\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right) p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} + \int \left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' \frac{\partial \ln p\left(\widetilde{\boldsymbol{\theta}}_3\right)}{\partial\boldsymbol{\theta}\boldsymbol{\theta}'}\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right) p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} \\
= & -\int \left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right)' L_n^{(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta} - \overleftrightarrow{\boldsymbol{\theta}}\right) p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} + o_p(1) + O_p(n^{-1}) \\
= & -\mathbf{tr}\left\{L_n^2\left(\overleftrightarrow{\boldsymbol{\theta}}\right) V\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right\} + o_p(1) \\
= & \mathbf{tr}\left\{L_n^{(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\left[L_n^{-(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) + o_p(n^{-1})\right]\right\} \\
= & \mathbf{tr}\left[L_n^{-(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) L_n^{-(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right] + \mathbf{tr}\left[L_n^{(2)}\left(\overleftrightarrow{\boldsymbol{\theta}}\right) o_p(n^{-1})\right] \\
= & P + o_p(1), \qquad (19)
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_3$ lies on the segment between $\boldsymbol{\theta}$ and $\overleftrightarrow{\boldsymbol{\theta}}$.

Finally, we have

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

by Assumption 10 and the dominated convergence theorem.

Proof of Theorem 4.1

Denote

$$\widetilde{\boldsymbol{\theta}} := \arg\max_{\boldsymbol{\theta}} \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right) + \ln p\left(\mathbf{y}|\boldsymbol{\theta}\right) + \ln p\left(\boldsymbol{\theta}\right).$$

Let

$$\ln p\left(\mathbf{y}|\boldsymbol{\theta}\right) = \sum_{t=1}^{n} l_t\left(y_t, \boldsymbol{\theta}\right), \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right) = \sum_{t=1}^{n} l_t\left(y_{t,rep}, \boldsymbol{\theta}\right),$$

and $\bigtriangledown l_t\left(y_t, \boldsymbol{\theta}\right)$ and $\bigtriangledown^2 l_t\left(y_t, \boldsymbol{\theta}\right)$ be the first and the second order derivatives of $l_t\left(y_t, \boldsymbol{\theta}\right)$. Then we have the following lemma under the condition that $\mathbf{y}$ and $\mathbf{y}_{rep}$ are independent:

Lemma 5.1 *Under Assumptions 1-15, if the prior of $\boldsymbol{\theta}$ is $O_p(1)$, $\widetilde{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}^t$.*

Proof. The proof follows the argument in Theorem 4.2 in Wooldridge (1994) and in Bester and Hansen (2006). Let $Q_n(\boldsymbol{\theta}) = n^{-1}\sum_{t=1}^{n} l_t(y_t, \boldsymbol{\theta}) + n^{-1}\sum_{t=1}^{n} l_t(y_{t,rep}, \boldsymbol{\theta}) + n^{-1}\ln p(\boldsymbol{\theta})$ and $\bar{Q}_n(\boldsymbol{\theta}) = E[Q_n(\boldsymbol{\theta})]$. Then we need to show that, for each $\varepsilon > 0$,

$$P\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| > \varepsilon\right] \to 0 \text{ as } n \to \infty.$$

Let $\delta > 0$ be a number to be set later. Because Θ is compact, there exists a finite number of spheres of radius δ about $\boldsymbol{\theta}_j$ say $\zeta_\delta(\boldsymbol{\theta}_j)$, $j = 1, \ldots, K(\delta)$, which cover Θ. Set $\zeta_j = \zeta_\delta(\boldsymbol{\theta}_j)$, $K = K(\delta)$. Because $\Theta \subset \cup_{j=1}^{K}\zeta_j$, it follows that

$$\begin{aligned} P\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| > \varepsilon\right] &\leq P\left[\max_{1\leq j\leq K}\sup_{\boldsymbol{\theta}\in\zeta_j}\left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| > \varepsilon\right] \\ &\leq \sum_{j=1}^{K} P\left[\sup_{\boldsymbol{\theta}\in\zeta_j}\left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| > \varepsilon\right]. \end{aligned}$$

For all $\boldsymbol{\theta} \in \zeta_j$,

$$\begin{aligned} \left|Q_n(\boldsymbol{\theta}) - \bar{Q}_n(\boldsymbol{\theta})\right| &\leq \left|Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}_j)\right| + \left|Q_n(\boldsymbol{\theta}_j) - \bar{Q}_n(\boldsymbol{\theta}_j)\right| + \left|\bar{Q}_n(\boldsymbol{\theta}_j) - \bar{Q}_n(\boldsymbol{\theta})\right| \\ &\leq \frac{1}{n}\sum_{t=1}^{n}\left|l_t(y_t, \boldsymbol{\theta}) - l_t(y_t, \boldsymbol{\theta}_j)\right| + \frac{1}{n}\sum_{t=1}^{n}\left|l_t(y_{t,rep}, \boldsymbol{\theta}) - l_t(y_{t,rep}, \boldsymbol{\theta}_j)\right| \\ &\quad + \left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(y_t, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right)\right| + \left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(y_{t,rep}, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right)\right| \\ &\quad + \frac{1}{n}\sum_{t=1}^{n}\left|\bar{l}_t(y_t, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right| + \frac{1}{n}\sum_{t=1}^{n}\left|\bar{l}_t(y_{t,rep}, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right| + \left|\frac{\ln p(\boldsymbol{\theta})}{n}\right|, \end{aligned}$$

where $\bar{l}_t(\boldsymbol{\theta}_j) := E[l_t(y_t, \boldsymbol{\theta})] = E[l_t(y_{t,rep}, \boldsymbol{\theta})]$. By Assumption 13, for all $\boldsymbol{\theta} \in \zeta_j$,

$$\left|l_t(y_t, \boldsymbol{\theta}) - l_t(y_t, \boldsymbol{\theta}_j)\right| \leq c_t(y_t)\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\right\| \leq \delta c_t(y_t).$$

and

$$\left|\bar{l}_t(y_t, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right| \leq E\left[c_t(y_t)\left\|\boldsymbol{\theta} - \boldsymbol{\theta}_j\right\|\right] \leq \delta\bar{c}_t,$$

where $\bar{c}_t = E[c_t(y_t)] = E[c_t(y_{t,rep})]$. Similarly we have

$$\left|l_t(y_{t,rep}, \boldsymbol{\theta}) - l_t(y_{t,rep}, \boldsymbol{\theta}_j)\right| \leq \delta c_t(y_{t,rep}), \ \left|\bar{l}_t(y_{t,rep}, \boldsymbol{\theta}) - \bar{l}_t(\boldsymbol{\theta}_j)\right| \leq \delta\bar{c}_t.$$

Thus, we have

$$\begin{aligned}
\sup_{\boldsymbol{\theta}\in\zeta_j}\left|Q_n\left(\boldsymbol{\theta}\right)-\bar{Q}_n\left(\boldsymbol{\theta}\right)\right| &\leq \frac{\delta}{n}\sum_{t=1}^{n}c_t\left(y_t\right)+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t \\
&\quad+\frac{\delta}{n}\sum_{t=1}^{n}c_t\left(y_{t,rep}\right)+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t \\
&\quad+\left|\frac{\ln p\left(\boldsymbol{\theta}\right)}{n}\right| \\
&\leq 2\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\
&\quad+2\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_{t,rep}\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\
&\quad+\left|\frac{\ln p\left(\boldsymbol{\theta}\right)}{n}\right| \\
&\leq 4\delta\bar{C}+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\
&\quad+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_{t,rep}\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\left|\frac{\ln p\left(\boldsymbol{\theta}\right)}{n}\right|,
\end{aligned}$$

where $n^{-1}\sum_{t=1}^{n}\bar{c}_t\leq\bar{C}<\infty$ by Assumption 13. It follows that

$$\begin{aligned}
&P\left[\max_{\boldsymbol{\theta}\in\zeta_j}\left|Q_n\left(\boldsymbol{\theta}\right)-\bar{Q}_n\left(\boldsymbol{\theta}\right)\right|>\varepsilon\right] \\
\leq\ &P\left[\begin{array}{c}\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\ +\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_{t,rep}\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\left|\frac{\ln p(\boldsymbol{\theta})}{n}\right|>\varepsilon-4\delta\bar{C}\end{array}\right].
\end{aligned}$$

Now choose $\delta\leq 1$ such that $\left(\varepsilon-2\delta\bar{C}\right)>\varepsilon/2$, then

$$\begin{aligned}
&P\left[\sup_{\boldsymbol{\theta}\in\zeta_j}\left|Q_n\left(\boldsymbol{\theta}\right)-\bar{Q}_n\left(\boldsymbol{\theta}\right)\right|>\varepsilon\right] \\
\leq\ &P\left[\begin{array}{c}\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\ +\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_{t,rep}\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\left|\frac{\ln p(\boldsymbol{\theta})}{n}\right|>\varepsilon/2\end{array}\right].
\end{aligned}$$

Next, choose n_0 so that

$$P\left[\begin{array}{c}\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_t,\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right| \\ +\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(y_{t,rep}\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(y_{t,rep},\boldsymbol{\theta}\right)-\bar{l}_t\left(\boldsymbol{\theta}_j\right)\right)\right|+\left|\frac{\ln p(\boldsymbol{\theta})}{n}\right|>\varepsilon/2\end{array}\right]\leq\frac{\varepsilon}{K}$$

for all $n\geq n_0$, and all $j=1,\ldots,K$ by Assumptions 5-15 since K is finite. Hence, for $n\geq n_0$

$$P\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|Q_n\left(\boldsymbol{\theta}\right)-\bar{Q}_n\left(\boldsymbol{\theta}\right)\right|>\varepsilon\right]\leq\varepsilon.$$

It then follows that $Q_n(\boldsymbol{\theta})$ satisfies a uniform weak law of large numbers and the consistency of $\widetilde{\boldsymbol{\theta}}$ followed by the usual argument. ■

Lemma 5.2 *Under Assumptions 1-15,* $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right) \xrightarrow{d} N\left(0,-\left(2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\right)$.

Proof. The proof follows from Bester and Hansen (2006). By Lemma 5.1, we have,

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(y_t,\widetilde{\boldsymbol{\theta}}\right)+\bigtriangledown l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}\right)\right]+\frac{1}{n}\ln p\left(\widetilde{\boldsymbol{\theta}}\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]+\frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_4\right)+\bigtriangledown^2 l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}_3\right)\right]\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right) \\
&\quad +\frac{\ln p\left(\widetilde{\boldsymbol{\theta}}\right)}{n}
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_4$ is an intermediate value between $\widetilde{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^t$. It follows that

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right) &= \left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_4\right)+\bigtriangledown^2 l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}_4\right)\right]\right)^{-1}\times \\
&\quad \left(n^{-1/2}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]+n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}\right)\right).
\end{aligned}$$

Under the assumptions, we have

$$n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}\right)=o_p(1),\quad -n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_4\right)\xrightarrow{p}-\mathbf{H}\left(\boldsymbol{\theta}^t\right),$$

$$-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_4\right)\xrightarrow{p}-\mathbf{H}\left(\boldsymbol{\theta}^t\right),-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}_4\right)\xrightarrow{p}-\mathbf{H}\left(\boldsymbol{\theta}^t\right),$$

$$n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)\xrightarrow{d}N\left(0,-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right),n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\xrightarrow{d}N\left(0,-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right).$$

Note that $\lim_{n\to\infty} Var\left(n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)\right)=-\mathbf{H}\left(\boldsymbol{\theta}^t\right)$. By the central limit theorem and the Cramer-Wold device, we get

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\xrightarrow{d}N\left(0,\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\right)\text{ or }\sqrt{2n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\xrightarrow{d}N\left(0,\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\right).$$

■

Lemma 5.3 *Under Assumption 1-15, the asymptotic joint distribution of* $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)$ *and* $\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)$ *is*

$$\begin{pmatrix}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\\ \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\end{pmatrix}\xrightarrow{d}N\left(\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1} & \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\\ \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1} & \left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\end{bmatrix}\right).$$

Proof. By Lemma 5.2, we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)=\left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_4\right)+\bigtriangledown^2 l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}_4\right)\right]\right)^{-1}$$
$$\times\left(n^{-1/2}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]+n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}\right)\right),$$

and

$$\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)=\left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_t,\widetilde{\boldsymbol{\theta}}_5\right)\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}\right)\right),$$

where $\widetilde{\boldsymbol{\theta}}_5$ is an intermediate value between $\overleftrightarrow{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}^t$. Hence, we have

$$\begin{aligned}
& Cov\left(\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right),\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\right)\\
= \;& E\left(\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\left[\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\right]'\right)\\
= \;& E\left[\begin{array}{c}\left\{-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(y_t,\boldsymbol{\theta}\right)+\bigtriangledown^2 l_t\left(y_{t,rep},\boldsymbol{\theta}\right)\right]\right\}^{-1}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)\\ \times\left[n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)\right]'\left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_t,\boldsymbol{\theta}_1\right)\right)^{-1}\end{array}\right]+o_p\left(1\right)\\
= \;& \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}+o_p\left(1\right)\\
= \;& \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}+o_p\left(1\right)
\end{aligned}$$

Then we have

$$\left(\begin{array}{c}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\\ \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\end{array}\right)\xrightarrow{d}N\left(\left[\begin{array}{c}0\\0\end{array}\right],\left[\begin{array}{cc}\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1} & \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\\ \left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1} & \left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1}\end{array}\right]\right).$$

■

Under Assumptions 1-15, it can be shown that,

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

By the Laplace approximation (Tierney et al., 1989 and Kass et al., 1990) and Lemma 5.2, we have

$$\begin{aligned}
p\left(\mathbf{y}_{rep}|\mathbf{y}\right) &= \int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}=\frac{\int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}}{\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}}\\
&= \frac{\int\exp\left(-nh_N\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}}{\int\exp\left(-nh_D\left(\boldsymbol{\theta}\right)\right)d\boldsymbol{\theta}}=\frac{\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right|^{-1/2}\exp\left(-nh_N\left(\widetilde{\boldsymbol{\theta}}\right)\right)}{\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|^{-1/2}\exp\left(-nh_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right)}\left(1+O_p\left(\frac{1}{n}\right)\right),
\end{aligned}$$

where

$$h_N\left(\boldsymbol{\theta}\right)=-\frac{1}{n}\left(\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)+\ln p\left(\mathbf{y}|\boldsymbol{\theta}\right)+\ln p\left(\boldsymbol{\theta}\right)\right),$$

$$h_D(\boldsymbol{\theta}) = -\frac{1}{n}\left(\ln p(\mathbf{y}|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})\right).$$

We know

$$\begin{aligned}
&\ln\left\{\frac{\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right|^{-1/2}\exp\left(-nh_N\left(\widetilde{\boldsymbol{\theta}}\right)\right)}{\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|^{-1/2}\exp\left(-nh_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right)}\right\} \\
= & -\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|\right) + \left[-nh_N\left(\widetilde{\boldsymbol{\theta}}\right) + nh_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right].
\end{aligned}$$

The first term is

$$\begin{aligned}
&-\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|\right) \\
= & -\frac{1}{2}\ln\left|-\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\widetilde{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \\
& +\frac{1}{2}\ln\left|-\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\overleftrightarrow{\boldsymbol{\theta}}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \\
= & -\frac{1}{2}\ln\left|-\mathbf{H}\left(\boldsymbol{\theta}^t\right) - \mathbf{H}\left(\boldsymbol{\theta}^t\right)\right| + \frac{1}{2}\ln\left|-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right| + o_p(1) \\
= & -\frac{1}{2}\ln\left|-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right| + \frac{1}{2}\ln\left|-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right| + o_p(1) \\
= & -\frac{1}{2}\ln\left(2^P\left|-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right|\right) + \frac{1}{2}\ln\left|-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right| + o_p(1) = -\frac{1}{2}P\ln 2 + o_p(1). \qquad (20)
\end{aligned}$$

Here we can see how $\ln 2$ shows up in the penalty term.

The second term is

$$\begin{aligned}
& -n\hat{h}_N\left(\widetilde{\boldsymbol{\theta}}\right) + n\hat{h}_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \\
= & \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) + \ln p\left(\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) - \ln p\left(\overleftrightarrow{\boldsymbol{\theta}}\right) \\
= & \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + o_p(1) \\
= & \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + o_p(1).
\end{aligned}$$

We can decompose $\ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right)$ as

$$\begin{aligned}
& \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) \\
= & \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right) + \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) := D_1 + D_2.
\end{aligned}$$

For D_1, we have

$$D_1 = \frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial\boldsymbol{\theta}}'\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right) + \frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right)'\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right) + o_p(1)$$

Following Assumption 5-15 and Lemma 5.3, we have

$$\begin{aligned}
& \frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)'}{\partial \boldsymbol{\theta}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right) \\
= & \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)'\left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)+o_p\left(1\right) \\
= & \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)'\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)+o_p\left(1\right) \\
= & \mathbf{tr}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)'\right]+o_p\left(1\right) \\
= & \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)'\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2} \\
& \times\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)+o_p\left(1\right) \\
= & 2^{-1/2}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right)+o_p\left(1\right).
\end{aligned}$$

where

$$\begin{aligned}
& \left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^t\right) \\
= & \left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(y_{t,rep},\widetilde{\boldsymbol{\theta}}_6\right)\right)^{-1}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right] \\
& \times\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{1/2}\left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(y_t,\boldsymbol{\theta}\right)+\bigtriangledown^2 l_t\left(y_{t,rep},\boldsymbol{\theta}\right)\right]\right)^{-1} \\
& \times\left[n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right] \\
= & \left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2} \\
& \times\left[n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right)+n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]+o_p\left(1\right) \\
= & \left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_t,\boldsymbol{\theta}^t\right) \\
& +2^{-1/2}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]'\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2} \\
& \times n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)+o_p\left(1\right),
\end{aligned}$$

with $\widetilde{\boldsymbol{\theta}}_6$ lying between $\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^t$, and

$$\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\right]'\left(-\mathbf{H}\left(\boldsymbol{\theta}^t\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(y_{t,rep},\boldsymbol{\theta}^t\right)\xrightarrow{d}\chi^2\left(P\right). \tag{21}$$

Thus, we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t,rep},\boldsymbol{\theta}^{t}\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t},\boldsymbol{\theta}^{t}\right)\right)=0, \tag{22}$$

since $\mathbf{y}$ and $\mathbf{y}_{rep}$ are independent. Hence, we have

$$\begin{aligned}
& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\frac{1}{\sqrt{n}}\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)'}{\partial\boldsymbol{\theta}}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)\right] \\
= & E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[2^{-1/2}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}\left(\mathbf{y}_{rep}\right)-\boldsymbol{\theta}^{t}\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)+o_{p}\left(1\right)\right] \\
= & E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(2^{-1/2}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t,rep},\boldsymbol{\theta}^{t}\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t},\boldsymbol{\theta}^{t}\right)\right) \\
& +E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\begin{array}{c}2^{-1/2}2^{-1/2}\left[\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t,rep},\boldsymbol{\theta}^{t}\right)\right]'\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2} \\ \times n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_{t}\left(y_{t,rep},\boldsymbol{\theta}^{t}\right)\end{array}\right]+o\left(1\right) \\
= & \frac{1}{2}P+o\left(1\right).
\end{aligned}$$

Moreover,

$$\begin{aligned}
& \frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)'\frac{1}{n}\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right) \\
= & -\frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)'\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)+o_{p}\left(1\right) \\
= & -\frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}\left(-\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{-1/2}\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2} \\
& \times\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)+o_{p}\left(1\right) \\
= & -\frac{1}{4}\left[\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)+o_{p}\left(1\right)
\end{aligned}$$

where

$$\left[\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)\right]'\left(-2\mathbf{H}\left(\boldsymbol{\theta}^{t}\right)\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)\overset{d}{\rightarrow}\chi^{2}\left(P\right). \tag{23}$$

For D_2, we have

$$\begin{aligned}
D_{2} = & \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)-\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)-\frac{1}{\sqrt{n}}\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)'}{\partial\boldsymbol{\theta}}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right) \\
& -\frac{1}{2}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)'\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)+o_{p}\left(1\right).
\end{aligned}$$

Since

$$-\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)'\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^{t}\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overleftrightarrow{\boldsymbol{\theta}}-\boldsymbol{\theta}^{t}\right)\overset{d}{\rightarrow}\chi^{2}\left(P\right), \tag{24}$$

$$E_{\mathbf{y}_{rep}}\left(\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)}{\partial \boldsymbol{\theta}}\right) = o\left(1\right), \tag{25}$$

from (21), (22), (23), (24), and (25), we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right) = \left(\frac{1}{2} - \frac{1}{4} + \frac{1}{2}\right)P. \tag{26}$$

Similarly, we can decompose $-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right)$ as

$$\begin{aligned} -\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) &= -\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right) \\ &= \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right) + \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right). \end{aligned}$$

From the discussion above, $\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right)$ has the same asymptotic property as $\ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}^t\right)$. Hence

$$\begin{aligned} & \ln p\left(\mathbf{y}|\boldsymbol{\theta}^t\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) \\ = \; & \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)'}{\partial \boldsymbol{\theta}}\sqrt{n}\left(\boldsymbol{\theta}^t - \overleftrightarrow{\boldsymbol{\theta}}\right) \\ & + \frac{1}{2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right)'\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + o_p\left(1\right), \end{aligned} \tag{27}$$

where

$$\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)'}{\partial \boldsymbol{\theta}} = o_p\left(1\right),$$

$$-\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right)'\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}} - \boldsymbol{\theta}^t\right) \xrightarrow{d} \chi^2\left(P\right).$$

Then

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right) = \left(\frac{1}{2} - \frac{1}{4} - \frac{1}{2}\right)P. \tag{28}$$

Note that

$$\bar{\boldsymbol{\theta}} = \overleftrightarrow{\boldsymbol{\theta}} + o_p(n^{-1/2}),$$

by Lemma 3.1. Mimicking the proof of Theorem 3.1, we get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) = E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] - P. \tag{29}$$

With (20), (26), (28) and (29), we have

$$
\begin{aligned}
& E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}} \ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right] \\
= \; & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \ln \left(\frac{\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right|^{-1/2} \exp\left(-n h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right)}{\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|^{-1/2} \exp\left(-n h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right)} \left(1 + O_p\left(\frac{1}{n}\right)\right)\right) \\
= \; & -\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}\right)\right|\right) \\
& + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] \\
& + o\left(1\right) \\
= \; & -\frac{P}{2}\ln 2 + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] \\
& + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] + o\left(1\right) \\
= \; & -\frac{P}{2}\ln 2 + E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right)\right] + \left(\frac{1}{2} - \frac{1}{4} + \frac{1}{2}\right)P + \left(\frac{1}{2} - \frac{1}{4} - \frac{1}{2}\right)P + o\left(1\right) \\
= \; & E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \left(\frac{1}{2} - \frac{\ln 2}{2}\right)P\right) + o\left(1\right) \\
= \; & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \left(\frac{1}{2} - \frac{\ln 2}{2}\right)P + o\left(1\right) \\
= \; & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \left(\frac{1}{2} - \frac{\ln 2}{2}\right)P + o\left(1\right) \\
= \; & E_{\mathbf{y}} \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) - P + \left(\frac{1}{2} - \frac{\ln 2}{2}\right)P + o\left(1\right) \\
= \; & E_{\mathbf{y}} E_{\mathbf{y}_{rep}} \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) + \left(-\frac{1}{2} - \frac{\ln 2}{2}\right)P + o\left(1\right) \\
= \; & E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}\right) - \frac{1+\ln 2}{2}P\right] + o\left(1\right), \\
= \; & E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) + o_p\left(1\right) - \frac{1+\ln 2}{2}P\right] + o\left(1\right) \\
= \; & E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) - \frac{1+\ln 2}{2}P\right] + o\left(1\right).
\end{aligned}
$$

Therefore, $-2\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}\right) + \left(1+\ln 2\right)P$ is an unbiased estimator of $E_{\mathbf{y}_{rep}}\left(-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)\right)$ asymptotically.

References

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

2 Andrews, D. W. K. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica*, **55(6)**, 1465-1471.

3 Andrews, D. W. K. (1988). Laws of large numbers for dependent non-identically distributed random variables. *Econometric theory*, **4**(03), 458-467.

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

5 Ando, T. (2012). Predictive Bayesian model selection. *American Journal of Mathematical and Management Sciences*, **31(1-2)**, 13-38.

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

7 Berger, J. O. (1985). *Statistical decision theory and Bayesian analysis*. 2nd edition. Springer-Verlag.

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

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

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

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

12 Gallant, A. R. and White, H. (1988). *A unified theory of estimation and inference for nonlinear dynamic models*. Blackwell.

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

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

15 Kass, R., Tierney, L. and Kadane, J. (1990) The validity of posterior expansions based on Laplace's Method. in *Bayesian and Likelihood Methods in Statistics and Econometrics*, ed. by S. Geisser, J.S. Hodges, S.J. Press and A. Zellner. Elsevier Science Publishers B.V.: North-Holland.

16 Jennrich, R. I. (1969). Asymptotic properties of non-linear least squares estimators. *The Annals of Mathematical Statistics*, **40(2)**, 633-643.

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

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

19 Phillips, P. C. B. (1995). Bayesian model selection and prediction with empirical application (with discussions). *Journal of Econometrics*, **69(1)**, 289-331.

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

21 Plummer, M. (2008). Penalized loss functions for Bayesian model comparison. *Biostatistics*, **9(3)**, 523-539.

22 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.

23 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.

24 Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential Laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84(407)**, 710-716.

25 van der Linde, A. (2005). DIC in variable selection. *Statistica Neerlandica*, **59(1)**, 45-56.

26 van der Linde, A. (2012). A Bayesian view of model complexity. *Statistica Neerlandica*, **66**, 253–271.

27 Vehtari, A., and Ojanen, J. (2012). A survey of Bayesian predictive methods for model assessment, selection and comparison. *Statistics Surveys*, **6**, 142-228.

28 Vehtari, A., and Lampinen, J. (2002). Bayesian model assessment and comparison using cross-validation predictive densities. *Neural Computation*, **14**, 2439-2468.

29 White, H. (1996). Estimation, inference and specification analysis. *Cambridge University Press*. Cambridge, UK.

30 Wooldridge, J. M. (1994). Estimation and inference for dependent processes. *Handbook of Econometrics*, **4**, 2639-2738.

31 Zhang, Y., and Yang, Y. (2015). Cross-validation for selecting a model selection procedure. *Journal of Econometrics*, **187(1)**, 95-112.