Deviance Information Criterion for Model Selection: Justification and Variation*

Yong Li
Renmin University of China

Jun Yu
Singapore Management University

Tao Zeng
Zhejiang University

November 23, 2017

Abstract

Deviance information criterion (DIC) has been extensively used for making model selection based on the MCMC output. Although it is understood as a Bayesian version of AIC, a rigorous justification has not been provided in the literature. In this paper, we show that when the plug-in predictive distribution is used, DIC can have a rigorous decision-theoretic justification in a frequentist setup. Under a set of regularity conditions, we show that DIC chooses a model that gives the smallest expected Kullback-Leibler divergence between the data generating process (DGP) and the plug-in predictive distribution asymptotically. 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 expected loss than the original DIC asymptotically.

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

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,

*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, China 100872. Jun Yu, School of Economics and Lee Kong Chian School of Business, Singapore Management University, 90 Stamford Rd, Singapore 178903. Email for Jun Yu: yujun@smu.edu.sg. URL: http://www.mysmu.edu/faculty/yujun/. Tao Zeng, School of Economics and Academy of Financial Research, Zhejiang University, Zhejiang, China 310027. Li gratefully acknowledges the financial support of the Chinese Natural Science Fund (No. 71271221), Program for New Century Excellent Talents in University.

the most popular information criterion is AIC. Arguably one of the most important 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 understood as 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 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 Jeffreys-Lindley's paradox and can be calculated when non-informative 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. In fact, in the heuristic justification given by Spiegelhalter, et al (2002), the frequentist framework and the Bayesian framework were mixed together. The first contribution of the present paper is to provide a rigorous decision-theoretic justification to DIC purely in a frequentist setup. 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 posterior mean is used. This justification is similar to how AIC has been justified.

Moreover, DIC is not invariant to reparameterization due to the use of the plug-in predictive distribution in defining the loss function. In the Bayesian framework, an alterative 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. The second contribution of the present paper is to develop a new information criterion that is an asymptotically unbiased estimator of the new expected KL divergence under a general framework. Compared with the information criterion developed in Ando and Tsay (2010) in the independent and identically distributed (iid) environment, our information criterion does not need the iid assumption and has a simpler expression. Moreover, it is easier to compare our information criterion with other information criteria.

Our theoretical results show that asymptotically the expected loss implied by the 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 predic-

[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 October 2017, it has received 4898 citations on the Web of Knowledge and over 8623 on Google Scholar.

tive 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 which helps explain our theory for DIC. Section 3 provides a rigorous decision-theoretic justification to DIC of Spiegelhalter, et al (2002) under a set of regularity conditions, and shows why DIC can be explained as the 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. To save space, the proof of Lemma 3.2 is provided in a technical supplement to the present article (Li et al., 2017).

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 maximum likelihood estimator. 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 is split into a training set 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. The latter approach is what we follow in the present paper. Given the relevance of prediction in economics, nor surprisingly, such an approach to model selection has been widely used in applications.

[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, $\mathbf{y} = (y_1, y_2, \cdots, y_n)' \in \mathbf{Y}$, 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 candidate model M_k and $p(\cdot)$ means a probability density function (pdf). 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, $\ell(\mathbf{y}, d_k)$, which measures the loss of decision d_k as a function of $\mathbf{y}$, must be specified. Given the loss function, the expected loss (or risk) can be defined as (Berger, 1985)

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

where $g(\mathbf{y})$ is the pdf of 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 $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)$ for M_k where $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ is the quasi maximum likelihood (QML) estimate of $\boldsymbol{\theta}_k$ based on $\mathbf{y}$ (when there is no confusion we simply write $\hat{\boldsymbol{\theta}}_n(\mathbf{y})$ as $\hat{\boldsymbol{\theta}}_n$ or even $\hat{\boldsymbol{\theta}}$) and defined by

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

which is the global maximum interior to $\boldsymbol{\Theta}$.

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 $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)$ multiplied by 2,

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

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

Naturally the loss function associated with decision d_k is

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

As a result, the model selection problem is,

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

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,

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

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

2.2 AIC for predictive model selection

To show the asymptotic unbiasedness of AIC, 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 $-2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)$ is a natural estimate of $E_{\mathbf{y}_{rep}}\left(-2\ln p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)$, it is asymptotically biased. Let

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

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

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

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

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

be the pseudo-true parameter value; $\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)$ be the QML estimate of $\boldsymbol{\theta}$ obtained from $\mathbf{y}_{rep}$; $p(\mathbf{y}|\boldsymbol{\theta})$ be a "good approximation" to the DGP, a concept will be defined formally in the next section. Note that

$$E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left(-2\ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right)$$

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

Clearly, the term in $T1$ is the same as $E_{\mathbf{y}}\left(-2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n(\mathbf{y})\right)\right)$. 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 $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ is a consistent estimate of $\boldsymbol{\theta}_n^p$, we have

$$
\begin{aligned}
T3 &= E_{\mathbf{y}}\left\{-2E_{\mathbf{y}_{rep}}\left[\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\theta'}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right]\right\} \\
&\quad + E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left[-\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\theta\partial\theta'}\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right]\right\} + o(1).
\end{aligned}
$$

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

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

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

$$
T3 = \mathbf{tr}\left\{E_{\mathbf{y}_{rep}}\left[\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\theta\partial\theta'}\right]E_{\mathbf{y}}\left[-\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\left(\widehat{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)'\right]\right\} = P + o(1),
$$

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

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

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

It is clear that the decision-theoretic justification of AIC rests on a frequentist framework. Specifically, it 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 $T3$ although a formal definition of "good approximation" was not provided.

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:

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

where $\bar{\boldsymbol{\theta}}_n(\mathbf{y})$ 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 $\bar{\boldsymbol{\theta}}_n(\mathbf{y})$ as $\bar{\boldsymbol{\theta}}_n$ or even $\bar{\boldsymbol{\theta}}$.

DIC can be rewritten in two equivalent forms:

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

and

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

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 $D\left(\bar{\boldsymbol{\theta}}_n\right) + P_D$ 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 stated explicitly 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 fact, it mixes a frequentist setup and a Bayesian setup. In this section, we provide a rigorous decision-theoretic justification of DIC purely in a frequentist setup. Specifically, we show that, when a proper loss function is selected, DIC is an asymptotically unbiased estimator of the expected loss.

3.2 Decision-theoretic justification of DIC

When developing DIC, Spiegelhalter, et al (2002) assumes that there is a true distribution for $\mathbf{y}$ in Section 2.2, a pseudo-true parameter value $\boldsymbol{\theta}_n^p$ for a candidate model also in Section 2.2, an independent replicate data set $\mathbf{y}_{rep}$ in Section 7.1. All these assumptions are identical what was done when justifying AIC. Furthermore, as explained in Section 7.1 of Spiegelhalter, et al (2002), the goal for model selection is to estimate the expected loss where the expectation is taken with respect to $\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p$. The assumptions and the goal indicate that a frequentist framework was considered. On the other hand, since the "optimism" associated with the natural estimator depends on a pseudo-true parameter value $\boldsymbol{\theta}_n^p$, instead of replacing it with a frequentist estimator and then finding the asymptotic property of the "optimism", Spiegelhalter, et al (2002) replaces $\boldsymbol{\theta}_n^p$ with a random quantity $\boldsymbol{\theta}$ and then calculates the posterior expectation of the "optimism"; see Sections 7.1 and 7.3. As a result, a Bayesian framework is adopted when studying the behavior of the "optimism".

Spiegelhalter, et al (2002) has not explicitly specified 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

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

was used.[3] Hence,

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

With this KL function, unfortunately, the first term in the right hand side of Equation (8) is no longer a constant across candidate models. This is because, when the pseudo-true value is replaced by a random quantity $\boldsymbol{\theta}$, the first term in the right hand side of Equation (8) is model dependent. Clearly, another KL divergence function 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

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

The corresponding expected loss function of a statistical decision d_k for model selection is

$$Risk(d_k) = E_{\mathbf{y}}\left\{E_{\mathbf{y}_{rep}}\left[2\ln \frac{g\left(\mathbf{y}_{rep}\right)}{p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_n(\mathbf{y}), M_k\right)}\right]\right\}$$

[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.

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

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 expected loss function $Risk(d_k)$ is equivalent to minimizing

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

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

We are now in the position to provide a rigorous decision-theoretic justification to DIC in a frequentist framework based on a set of regularity conditions. Let $\mathbf{y}^t := (y_0, y_1, \ldots, y_t)$ for any $0 \leq t \leq n$ and $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right) := \ln p(\mathbf{y}^t | \boldsymbol{\theta}) - \ln p(\mathbf{y}^{t-1} | \boldsymbol{\theta})$ be the conditional log-likelihood for the t^{th} observation for any $1 \leq t \leq n$. When there is no confusion, we suppress $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right)$ as $l_t(\boldsymbol{\theta})$ so that the log-likelihood function $\ln p(\mathbf{y} | \boldsymbol{\theta})$ is $\sum_{t=1}^n l_t(\boldsymbol{\theta})$.[4] Let $\nabla^j l_t(\boldsymbol{\theta})$ denote the j^{th} derivative of $l_t(\boldsymbol{\theta})$ and $\nabla^j l_t(\boldsymbol{\theta}) := l_t(\boldsymbol{\theta})$ when $j = 0$. Furthermore, define

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

In this paper, we impose the following regularity conditions.

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

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

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

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

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

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

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

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

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

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

$$\lim_{n\to\infty}\sup \sup_{\boldsymbol{\Theta}\setminus N(\boldsymbol{\theta}_n^p,\varepsilon)} \frac{1}{n}\sum_{t=1}^{n}\left\{E\left[l_t(\boldsymbol{\theta})\right] - E\left[l_t(\boldsymbol{\theta}_n^p)\right]\right\} < 0, \tag{9}$$

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

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

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

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

Remark 3.1 *Assumption 1 is the compactness condition. Assumption 2 and Assumption 6 imply weak dependence in y_t and l_t. The first part of Assumption 3 is the continuity condition. Assumption 4 is the Lipschitz condition for l_t first introduced in Andrews (1987) to develop the uniform law of large numbers for dependent and heterogeneous stochastic processes. Assumption 5 contains the domination condition for l_t. Assumption 7 is the identification condition. These assumptions are well-known primitive conditions for developing the QML theory, namely consistency and asymptotic normality, for dependent and heterogeneous data; see, for example, Gallant and White (1988) and Wooldridge (1994).*

Remark 3.2 *The eight-times differentiability condition in Assumption 3 and the domination condition for up to the eighth derivative of l_t are important to develop a high order stochastic Laplace expansion. In particular, as shown in Kass et al (1990), these two conditions, together with the well-known consistency condition for QML given by (10) below, are sufficient for developing the Laplace expansion. This consistency condition requires that, for any $\varepsilon > 0$, there exists $K_1(\varepsilon) > 0$ such that*

$$\lim_{n\to\infty} P\left(\sup_{\boldsymbol{\Theta}\setminus N(\boldsymbol{\theta}_n^p,\varepsilon)} \frac{1}{n}\sum_{t=1}^{n}\left[l_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}_n^p)\right] < -K_1(\varepsilon)\right) = 1. \tag{10}$$

Our Assumption 7 is clearly more primitive than the consistency condition (10). In the following lemma, we show that Assumptions 1-7, including the identification condition (9), are sufficient to ensure (10) as well as the concentration condition around the posterior mode

given by Chen (1985) and the concentration condition around the QML estimator given by Kim (1994, 1998). Together with Assumption 10, the concentration condition suggests that the stochastic Laplace expansion can be applied to the posterior distribution and the asymptotic normality of posterior distribution can be established. To the best of our knowledge, this is the first time in the literature that primitive conditions have been proposed for the stochastic Laplace expansion. Assumption 10 ensures the second moment of the prior is bounded. As argued in Geweke (2001), such a condition typically leads to a finite second moment of posterior. Moreover, it implies that the prior is negligible asymptotically.

Lemma 3.1 *If Assumptions 1-7 hold true, then Equation (10) holds. Furthermore, if Assumptions 1-7 hold true, for any $\varepsilon > 0$, there exists $K_2(\varepsilon) > 0$ such that*

$$\lim_{n\to\infty} P\left(\sup_{\boldsymbol{\Theta}\setminus N\left(\widehat{\boldsymbol{\theta}}_{n,\varepsilon}\right)} \frac{1}{n}\left[\sum_{t=1}^{n} l_t(\boldsymbol{\theta}) - \sum_{t=1}^{n} l_t(\boldsymbol{\theta}_n^p)\right] < -K_2(\varepsilon)\right) = 1. \tag{11}$$

If, in addition, Assumption 10 holds true, for any $\varepsilon > 0$, there exists $K_3(\varepsilon) > 0$ such that

$$\lim_{n\to\infty} P\left(\sup_{\boldsymbol{\Theta}\setminus N\left(\overleftrightarrow{\boldsymbol{\theta}}_{n,\varepsilon}\right)} \frac{1}{n}\left(\sum_{t=1}^{n}\left[l_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}_n^p)\right] + \ln p(\boldsymbol{\theta}) - \ln p(\boldsymbol{\theta}_n^p)\right) < -K_3(\varepsilon)\right) = 1, \tag{12}$$

where $\overleftrightarrow{\boldsymbol{\theta}}_n := \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \sum_{t=1}^{n} l_t(\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$ which is the posterior mode.

Remark 3.3 *Assumption 9 gives the exact requirement for "good approximation". This generalizes the definition of "information matrix equality"; see White (1996). We now specify a few cases where Assumption 9 is satisfied. The detail of these cases can be found in White (1996) and Wooldridge (1994).*

1. *According to White (1996, p55), a model is specification correct in its entirety if there exists $\boldsymbol{\theta}_0 \in \Theta$ such that*

$$p(\mathbf{y}|\boldsymbol{\theta}_0) = g(\mathbf{y}). \tag{13}$$

   In this case, $\boldsymbol{\theta}_n^p = \boldsymbol{\theta}_0$ for all n, $\mathbf{H}_n(\boldsymbol{\theta}_0) + \mathbf{B}_n(\boldsymbol{\theta}_0) = 0$, implying Assumption 9 (White, 1996, p93).

2. *For any t, if $Cov\left[\mathbf{s}_t(\boldsymbol{\theta}_n^p), \mathbf{s}_{t+j}(\boldsymbol{\theta}_n^p)\right] = 0$ for all $j \geqslant 1$, then $\mathbf{J}_n(\boldsymbol{\theta}_n^p) = \mathbf{B}_n(\boldsymbol{\theta}_n^p)$. For any t, if $Var\left[\mathbf{s}_t(\boldsymbol{\theta}_n^p)\right] = -E\left[\mathbf{h}_t(\boldsymbol{\theta}_n^p)\right]$, then $\mathbf{J}_n(\boldsymbol{\theta}_n^p) + \mathbf{H}_n(\boldsymbol{\theta}_n^p) = 0$. If both conditions are satisfied, we have $\mathbf{H}_n(\boldsymbol{\theta}_n^p) + \mathbf{B}_n(\boldsymbol{\theta}_n^p) = 0$, again implying Assumption 9.*

3. *Suppose y_t is a $k_y \times 1$ vector which can be partitioned into a $k_w \times 1$ vector of endogenous variables w_t and a $k_z \times 1$ vector of exogenous variables z_t. Let x_t be a subset*

of $(z_t, w_{t-1}, z_{t-1}, \ldots, w_1, z_1)$. *According to White (1996, p55), a model is specification correct for dependent variable* w_t, *if there exists* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, *such that*

$$p\left(w_t|x_t, \boldsymbol{\theta}_0\right) = g\left(w_t|x_t\right), \text{ all } x_t, \; t = 1, 2, \ldots. \tag{14}$$

If $l_t\left(\boldsymbol{\theta}\right) = \ln p\left(w_t|x_t, \boldsymbol{\theta}\right)$, *then* $E\left[\mathbf{s}_t(\boldsymbol{\theta}_0)|x_t\right] = 0$. *Under the condition (14),* $\boldsymbol{\theta}_n^p = \boldsymbol{\theta}_0$ *for all* n, $E\left[\mathbf{s}_t(\boldsymbol{\theta}_0)\mathbf{s}_t(\boldsymbol{\theta}_0)'\right] = -E\left[\mathbf{h}_t(\boldsymbol{\theta}_0)\right]$, *and* $J_n\left(\boldsymbol{\theta}_0\right) = -H_n\left(\boldsymbol{\theta}_0\right)$. *If the model is further assumed to be dynamically complete in distribution in the sense that*

$$g\left(w_t|x_t, w_{t-1}, z_{t-1}, \ldots, w_1, z_1\right) = g\left(w_t|x_t\right),$$

then $E\left[\mathbf{s}_t(\boldsymbol{\theta}_0)\mathbf{s}_{t+j}(\boldsymbol{\theta}_0)'\right] = 0$ *and* $\mathbf{J}_n\left(\boldsymbol{\theta}_0\right) = \mathbf{B}_n\left(\boldsymbol{\theta}_0\right)$. *Hence,* $\mathbf{H}_n\left(\boldsymbol{\theta}_0\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_0\right) = 0$, *implying Assumption 9; see White (1996, p95-97).*

4. *A model is specification correct for the conditional mean and the conditional variance, if there exists* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$, *such that*

$$E\left(w_t|x_t\right) = m_t\left(x_t, \boldsymbol{\theta}_0\right), Var\left(w_t|x_t\right) = \Omega_t\left(x_t, \boldsymbol{\theta}_0\right), \text{for any } t,$$

where $m_t\left(x_t, \boldsymbol{\theta}\right)$ *and* $\Omega_t\left(x_t, \boldsymbol{\theta}\right)$ *are the conditional mean and the conditional variance of the model. In this case, the QML method that maximizes the Gaussian-quasi-log-likelihood function can be used. Let*

$$l_t\left(\boldsymbol{\theta}\right) := -\frac{1}{2}\ln\left|\Omega_t\left(x_t, \boldsymbol{\theta}\right)\right| - \frac{1}{2}u_t\Omega_t\left(x_t, \boldsymbol{\theta}\right)^{-1}u_t,$$

where $u_t = u_t\left(\boldsymbol{\theta}\right) = w_t - m_t\left(x_t, \boldsymbol{\theta}\right)$. *It is easy to show that* $E\left[\mathbf{s}_t(\boldsymbol{\theta}_n^p)|x_t\right] = 0$. *If the model is further assumed to be dynamically complete in mean and variance in the sense that*

$$E\left(w_t|x_t, w_{t-1}, z_{t-1}, \ldots, w_1, z_1\right) = E\left(w_t|x_t\right), Var\left(w_t|x_t, w_{t-1}, z_{t-1}, \ldots, w_1, z_1\right) = Var\left(w_t|x_t\right), \tag{15}$$

then $\boldsymbol{\theta}_n^p = \boldsymbol{\theta}_0$, $\left\{\mathbf{s}_t(\boldsymbol{\theta}_0)\right\}$ *is a martingale difference sequence, and* $\mathbf{J}_n\left(\boldsymbol{\theta}_0\right) = \mathbf{B}_n\left(\boldsymbol{\theta}_0\right)$. *If*

$$E\left[vec(u_tu_t')u_t'|x_t\right] = 0,$$

$$E\left[\left\{vec(u_tu_t') - \Omega_t\left(x_t, \boldsymbol{\theta}_0\right)\right\}\left\{vec(u_tu_t') - \Omega_t\left(x_t, \boldsymbol{\theta}_0\right)\right\}'|x_t\right] = 2N_{k_w}\left[\Omega_t\left(x_t, \boldsymbol{\theta}_0\right) \otimes \Omega_t\left(x_t, \boldsymbol{\theta}_0\right)\right],$$

where vec is the column-wise vectorization, $N_{k_w} = D_{k_w}\left(D_{k_w}'D_{k_w}\right)^{-1}D_{k_w}'$, *and* D_{k_w} *is the* $k_w^2 \times k_w(k_w+1)/2$ *duplication matrix, then* $\mathbf{J}_n\left(\boldsymbol{\theta}_0\right) = -\mathbf{H}_n\left(\boldsymbol{\theta}_0\right)$. *Hence,* $\mathbf{H}_n\left(\boldsymbol{\theta}_0\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_0\right) = 0$, *implying Assumption 9; see Wooldridge (1994, p2690).*

The above four cases all imply that $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) = 0$, stronger than what Assumption 9 requires. We now give an example where $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) = o\left(1\right)$.

Example 3.1 *Let the DGP be*

$$y_t = x_{1t}\beta_0 + x_{2t}\gamma_0 + \varepsilon_t,\ \varepsilon_t \overset{iid}{\sim} N(0, \sigma_0^2),$$

where (x_{1t}, x_{2t}) *is iid over* t *and independent of* ε_t*. Assume that* $\gamma_0 = \delta_0/n^{1/2}$*, where* δ_0 *is an unknown constant. Let the candidate model be*

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

In this case

$$l_t(\boldsymbol{\theta}) = -\frac{1}{2}\ln 2\pi - \frac{1}{2}\ln\sigma^2 - \frac{(y_t - x_{1t}\beta)^2}{2\sigma^2},$$

where $\boldsymbol{\theta} = \left(\beta, \sigma^2\right)'$*. In this case, the pseudo true value is* $\boldsymbol{\theta}_n^p = \left(\beta_n^p, \sigma_n^{2p}\right)'$*, which maximizes* $E\left[l_t(\boldsymbol{\theta})\right]$*, and can be expressed as*

$$\beta_n^p = \beta_0 + b\gamma_0, \sigma_n^{2p} = \sigma_0^2 + c\gamma_0^2,$$

where $b = \left[E\left(x_{1t}^2\right)\right]^{-1} E\left(x_{1t}x_{2t}\right)$ *and* $c = E\left(x_{2t}^2\right) - \left[E\left(x_{1t}x_{2t}\right)\right]^2 \left[E\left(x_{1t}^2\right)\right]^{-1}$*. Hence,*

$$-E\left[\mathbf{h}_t\left(\boldsymbol{\theta}_n^p\right)\right] = \begin{bmatrix} \frac{E\left(x_{1t}^2\right)}{\sigma_n^{2p}} & 0 \\ 0 & -\frac{1}{2\left(\sigma_n^{2p}\right)^2} - \frac{\sigma_0^2 + c\gamma_0^2}{\left(\sigma_n^{2p}\right)^3} \end{bmatrix},$$

$$-\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) = -\frac{1}{n}\sum_{t=1}^{n} E\left[\mathbf{h}_t\left(\boldsymbol{\theta}_n^p\right)\right] = -E\left[\mathbf{h}_t\left(\boldsymbol{\theta}_n^p\right)\right].$$

From the iid assumption, we have

$$\begin{aligned} & Var\left(s_t\left(\boldsymbol{\theta}_n^p\right)\right) = E\left(s_t\left(\boldsymbol{\theta}_n^p\right) s_t\left(\boldsymbol{\theta}_n^p\right)'\right) \\ = & \begin{bmatrix} \frac{\sigma_0^2 E\left(x_{1t}x_{1t}'\right)}{\sigma_n^{2p}} + \frac{d_1\gamma_0^2}{\left(\sigma_n^{2p}\right)^2} & \frac{d_2\gamma_0^3}{2\left(\sigma_n^{2p}\right)^2} \\ \frac{d_2\gamma_0^3}{2\left(\sigma_n^{2p}\right)^2} & -\frac{1}{4\left(\sigma_n^{2p}\right)^2} + \frac{3\sigma_0^2 + 6c\sigma_0^2\gamma_0^2 + d_3\gamma_0^4}{4\left(\sigma_n^{2p}\right)^4} \end{bmatrix}. \end{aligned}$$

where $d_j = E\left[x_{1t}^{4-j-1}\left(x_{2t} - x_{1t}b\right)^{j+1}\right]$ *for* $j = 1, 2, 3$ *and*

$$\begin{aligned} \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) &= Var\left(\frac{1}{\sqrt{n}}\sum_{t=1}^{n} s_t\left(\boldsymbol{\theta}_n^p\right)\right) = \frac{1}{n}\sum_{t=1}^{n} Var\left(s_t\left(\boldsymbol{\theta}_n^p\right)\right) = \mathbf{J}_n\left(\boldsymbol{\theta}_n^p\right) \\ &= \begin{bmatrix} \frac{\sigma_0^2 E\left(x_{1t}^2\right)}{\sigma_n^{2p}} + \frac{d_1\gamma_0^2}{\left(\sigma_n^{2p}\right)^2} & \frac{d_2\gamma_0^3}{2\left(\sigma_n^{2p}\right)^2} \\ \frac{d_2\gamma_0^3}{2\left(\sigma_n^{2p}\right)^2} & -\frac{1}{4\left(\sigma_n^{2p}\right)^2} + \frac{3\left(\sigma_0^2\right)^2 + 6c\sigma_0^2\gamma_0^2 + d_3\gamma_0^4}{4\left(\sigma_n^{2p}\right)^4} \end{bmatrix}. \end{aligned}$$

Hence,

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

since $\gamma_0 = \delta_0/n^{1/2}$*. Thus,* $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) = o(1)$*. However,* $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) \neq 0$ *for any finite* n*.*

To develop the Laplace expansion, we need to fix more notations. For convenience of exposition, we let $\bar{\mathbf{H}}_n^{(j)}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} \nabla^j l_t(\boldsymbol{\theta})$ for $j = 3, 4, 5$. Let $\pi(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta})$, $\widehat{p}$, $\widehat{\pi}$, $\nabla^j \widehat{p}$, and $\nabla^j \widehat{\pi}$ be the values of functions, $p(\boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$, $\nabla^j p(\boldsymbol{\theta})$ and $\nabla^j \pi(\boldsymbol{\theta})$ evaluated at $\widehat{\boldsymbol{\theta}}_n$. The next lemma extends the result in Kass et al (1990) to a higher order in matrix form.

Lemma 3.2 *Under Assumptions 1-10, we have*

$$\frac{\int l_t(\boldsymbol{\theta}) p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}}{\int p(\boldsymbol{\theta}) p(\mathbf{y}|\boldsymbol{\theta}) d\boldsymbol{\theta}} = l_t\left(\widehat{\boldsymbol{\theta}}_n\right) + \frac{1}{n} B_{t,1} + \frac{1}{n^2}\left(B_{t,2} - B_{t,3}\right) + O_p\left(n^{-3}\right), \tag{16}$$

where

$$\begin{aligned} B_{t,1} &= -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] - \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla \widehat{p}}{\widehat{p}} \\ &\quad -\frac{1}{2} vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right) \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right), \end{aligned} \tag{17}$$

$$\begin{aligned} B_{t,2} &= -\frac{1}{8} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bar{\mathbf{H}}_n^{(5)}\left(\widehat{\boldsymbol{\theta}}_n\right)' vec\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\right] \\ &\quad +\frac{35}{48} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' A_2 \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \\ &\quad -\frac{35}{48} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) A_1 \\ &\quad -\frac{5}{8} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla \widehat{p}}{\widehat{p}} \mathbf{tr}\left[A_2\right] + \frac{35}{24} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla \widehat{p}}{\widehat{p}} A_1 \\ &\quad -\frac{5}{4} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla^2 \widehat{p}}{\widehat{p}}\right] \\ &\quad +\frac{1}{2} \bigtriangledown l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)' \frac{\nabla^3 \widehat{p}}{\widehat{p}} \left[vec\left(\bar{\mathbf{H}}_n^{(2)}\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\right] \\ &\quad -\frac{5}{16}\mathbf{tr}\left[A_2\right]\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] + \frac{35}{48} A_1 \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] \\ &\quad -\frac{5}{12} vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)' \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right) \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)' vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \\ &\quad +\frac{1}{8}\mathbf{tr}\left(\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\right]' \bigtriangledown^4 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right) \\ &\quad -\frac{5}{4} vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)' \bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right) \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \bigtriangledown^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right) vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \frac{\nabla \widehat{p}}{\widehat{p}} \\ &\quad +\frac{1}{2} vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)' \bigtriangledown^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right) \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla \widehat{p}}{\widehat{p}} \\ &\quad +\frac{3}{4}\mathbf{tr}\left[\bigtriangledown^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right) \bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right] \mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \frac{\nabla^2 \widehat{p}}{\widehat{p}}\right], \end{aligned} \tag{18}$$

$$B_{t,3} = B_4 \times B_{t,1} \text{ with} \tag{19}$$

$$B_4 = -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla^2\widehat{p}}{\widehat{p}}\right] + \frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}}$$
$$-\frac{5}{24}A_1 + \frac{1}{8}\mathbf{tr}\left[A_2\right], \tag{20}$$

$$A_1 = vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right),$$

$$A_2 = \left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\right]'\bar{\mathbf{H}}_n^{(4)}\left(\widehat{\boldsymbol{\theta}}_n\right).$$

We are now in the position to develop a high order expansion to P_D and DIC.

Lemma 3.3 *Under Assumptions 1-10, we have*

$$P_D = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(n^{-2}\right),$$
$$DIC = AIC + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(n^{-2}\right),$$

where

$$C_1 = -\frac{-13+15P}{12}A_1 - \frac{1-2P}{4}\mathbf{tr}\left[A_2\right], C_2 = \frac{3-P}{2}C_{21} - PC_{22} + (1-P)C_{23},$$

$$D_1 = -\frac{1-2P}{2}\mathbf{tr}\left[A_2\right] - \frac{-23+25P}{12}A_1, D_2 = (2-P)C_{21} - 2PC_{22} + (1-2P)C_{23},$$

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

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

where A_1 and A_2 are defined as in Lemma 3.2.

Theorem 3.1 *Under Assumptions 1-10, we have,*

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

Remark 3.4 *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 asymptotically minimizes the expected loss, which is the expected KL divergence between the DGP and the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_n(\mathbf{y})\right)$ 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, $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$, is used while in DIC the Bayesian posterior mean, $\bar{\boldsymbol{\theta}}_n(\mathbf{y})$, is used. That is why DIC is explained as the Bayesian version of AIC.*

Remark 3.5 *The justification of DIC remains valid if the posterior mean is 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).*

Remark 3.6 *In AIC, the number of degrees of freedom, P, is used to measure the model complexity. When the prior is informative, it imposes additional restrictions on the parameter space and, hence, 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). From Lemma 3.3, the effect of prior information on P_D is reflected by C_2 and is of order $O_p\left(n^{-1}\right)$. Similarly, the effect of prior information on DIC is reflected by D_2 and is also of order $O_p\left(n^{-1}\right)$.*

Remark 3.7 *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 3.8 *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, are among the reasons why DIC has enjoyed 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 Tsay (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

$$KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\mathbf{y},M_k\right)\right] = E_{\mathbf{y}_{rep}}\left(\ln g\left(\mathbf{y}_{rep}\right)\right) - E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\mathbf{y},M_k\right)\right). \tag{21}$$

The expected loss for a statistical decision d_k that selects Model M_k is

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

A better model is expected to yield a smaller value for $Risk(d_k)$. 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 (21) 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 Tsay (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

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

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

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 Tsay (2010) defined the following information criterion*

$$IC_{AT} = -2\ln p(\mathbf{y}|\mathbf{y}) + \mathbf{tr}\left[\mathbf{I}_{AT}^{-1}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\mathbf{J}_{AT}\left(\hat{\boldsymbol{\theta}}_{AT}\right)\right]. \tag{22}$$

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 method 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 Tsay simplified the information criterion as*

$$IC_{AT} = -2\ln p(\mathbf{y}|\mathbf{y}) + P. \tag{23}$$

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., $D\left(\bar{\boldsymbol{\theta}}\right)$ but without resorting the iid assumption. It turns out such a choice not only facilitates the comparison with DIC, AIC or BIC, but also leads to a much simpler expression for the penalty term. 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 as*

$$DIC^{BP} = D\left(\bar{\boldsymbol{\theta}}\right) + (1+\ln 2)P_D, \tag{24}$$

where P_D is defined in (4). Under Assumptions 1-10, 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 1-10, $p(\boldsymbol{\theta}|\mathbf{y})$ 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. Hence, the prior information is incorporated in* DIC^{BP}*.*

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

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

Under the iid assumption, it was shown that

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

leading to an information criterion called BPIC by Ando (2007). Clearly, the target here is different from that under DIC and also from that under DIC^{BP}*. According to Spiegelhalter, et al (2014) and Vehtari and Ojanen (2012), BPIC chooses an "average target" rather than a "representative target". Although BPIC can select a model, it cannot tell the user how to actually predict the future observations because there is no utility corresponding to this criterion.*

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 expected loss in DIC^{BP} is smaller than that in BPIC for the same candidate model.

4.3 Expected loss 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 expected loss 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

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

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 expected loss associated with d_{k^a} is

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

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

$$\min_{a\in\{0,1\}} \min_{k\in\{1,\cdots,K\}} Risk\left(d_{k^a}\right). \tag{25}$$

According to Lemma 3.3, $P_D = P + o_p(1)$. 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

$$Risk\left(d_{k^0}\right) = E_{\mathbf{y}}\left(\ell\left(\mathbf{y}, d_{k^0}\right)\right) > E_{\mathbf{y}}\left(\ell\left(\mathbf{y}, d_{k^1}\right)\right) = Risk\left(d_{k^1}\right). \tag{26}$$

Hence, w.p.a.1,

$$\min_{k\in\{1,\cdots,K\}} Risk(d_{k^0}) > \min_{k\in\{1,\cdots,K\}} Risk\left(d_{k^1}\right). \tag{27}$$

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 expected loss 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 expected loss 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 expected loss 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	$\overleftrightarrow{\boldsymbol{\theta}}_n$	posterior mode
$o(1)$	tend to zero	$\widehat{\boldsymbol{\theta}}_n$	QML estimate
$o_p(1)$	tend to zero in probability	$\boldsymbol{\theta}_n^p$	pseudo true parameter
$\xrightarrow{p}$	converge in probability	$\hat{\boldsymbol{\theta}}_{AT}$	$\arg\max$ of $2\ln p(\mathbf{y}\|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$
$\overline{\boldsymbol{\theta}}_n$	posterior mean	$\widetilde{\boldsymbol{\theta}}_n$	$\arg\max$ of $\ln p(\mathbf{y}_{rep}\|\boldsymbol{\theta}) + \ln p(\mathbf{y}\|\boldsymbol{\theta}) + \ln p(\boldsymbol{\theta})$

Proof of Lemma 3.1

$$\begin{aligned} & \frac{1}{n}\sum_{t=1}^{n}\left[l_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}_n^p)\right] \\ = \; & \frac{1}{n}\sum_{t=1}^{n}\left(l_t(\boldsymbol{\theta}) - E\left[l_t(\boldsymbol{\theta})\right]\right) + \frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t(\boldsymbol{\theta})\right] - E\left[l_t(\boldsymbol{\theta}_n^p)\right]\right) + \frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t(\boldsymbol{\theta}_n^p)\right] - l_t(\boldsymbol{\theta}_n^p)\right). \end{aligned}$$

From (9), we know that for any $\varepsilon > 0$, there exists $\delta_1(\varepsilon) > 0$ and $N(\varepsilon) > 0$, for all $n > N(\varepsilon)$,

$$\frac{1}{n}\sum_{t=1}^{n}\left\{E\left[l_t(\boldsymbol{\theta})\right] - E\left[l_t(\boldsymbol{\theta}_n^p)\right]\right\} < -\delta_1(\varepsilon),$$

if $\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash N(\boldsymbol{\theta}_n^p, \varepsilon)$. Thus, for any $\varepsilon > 0$, if $\boldsymbol{\theta} \in \boldsymbol{\Theta}\backslash N(\boldsymbol{\theta}_n^p, \varepsilon)$, for all $n > N(\varepsilon)$,

$$\frac{1}{n}\sum_{t=1}^{n}\left[l_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}_n^p)\right] < \frac{1}{n}\sum_{t=1}^{n}\left(l_t(\boldsymbol{\theta}) - E\left[l_t(\boldsymbol{\theta})\right]\right) - \delta_1(\varepsilon) + \frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t(\boldsymbol{\theta}_n^p)\right] - l_t(\boldsymbol{\theta}_n^p)\right),$$

and

$$\begin{aligned} & \sup_{\boldsymbol{\Theta}\backslash N(\boldsymbol{\theta}_n^p,\varepsilon)} \frac{1}{n}\sum_{t=1}^{n}\left[l_t(\boldsymbol{\theta}) - l_t(\boldsymbol{\theta}_n^p)\right] \\ \leq \; & \sup_{\boldsymbol{\Theta}\backslash N(\boldsymbol{\theta}_n^p,\varepsilon)} \frac{1}{n}\sum_{t=1}^{n}\left(l_t(\boldsymbol{\theta}) - E\left[l_t(\boldsymbol{\theta})\right]\right) - \delta_1(\varepsilon) + \frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t(\boldsymbol{\theta}_n^p)\right] - l_t(\boldsymbol{\theta}_n^p)\right) \\ \leq \; & \sup_{\boldsymbol{\Theta}\backslash N(\boldsymbol{\theta}_n^p,\varepsilon)} \left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t(\boldsymbol{\theta}) - E\left[l_t(\boldsymbol{\theta})\right]\right)\right| - \delta_1(\varepsilon) + \left|\frac{1}{n}\sum_{t=1}^{n}\left(E\left[l_t(\boldsymbol{\theta}_n^p)\right] - l_t(\boldsymbol{\theta}_n^p)\right)\right| \end{aligned}$$

$$\leq \quad 2 \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(l_t \left(\boldsymbol{\theta} \right) - E \left[l_t \left(\boldsymbol{\theta} \right) \right] \right) \right| - \delta_1 \left(\varepsilon \right). \tag{28}$$

Under Assumptions 1-6, the uniform convergence condition is satisfied, i.e.,

$$P \left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(l_t \left(\boldsymbol{\theta} \right) - E \left[l_t \left(\boldsymbol{\theta} \right) \right] \right) \right| < \varepsilon \right) \to 1, \tag{29}$$

From the uniform convergence, if we choose δ_2 such that $0 < \delta_2 < \delta_1 \left(\varepsilon \right) / 2$, we have

$$P \left[\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(l_t \left(\boldsymbol{\theta} \right) - E \left[l_t \left(\boldsymbol{\theta} \right) \right] \right) \right| < \delta_2 \right] \to 1.$$

Hence,

$$P \left[2 \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(l_t \left(\boldsymbol{\theta} \right) - E \left[l_t \left(\boldsymbol{\theta} \right) \right] \right) \right| - \delta_1 \left(\varepsilon \right) < 2\delta_2 - \delta_1 \left(\varepsilon \right) \right] \to 1.$$

From (28), we have

$$\begin{aligned} & P \left[2 \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left| \frac{1}{n} \sum_{t=1}^{n} \left(l_t \left(\boldsymbol{\theta} \right) - E \left[l_t \left(\boldsymbol{\theta} \right) \right] \right) \right| - \delta_1 \left(\varepsilon \right) < 2\delta_2 - \delta_1 \left(\varepsilon \right) \right] \\ \leqslant \quad & P \left[\sup_{\boldsymbol{\Theta}\setminus N\left(\boldsymbol{\theta}_n^p, \varepsilon \right)} \frac{1}{n} \left[\sum_{t=1}^{n} l_t \left(\boldsymbol{\theta} \right) - \sum_{t=1}^{n} l_t \left(\boldsymbol{\theta}_n^p \right) \right] < 2\delta_2 - \delta_1 \left(\varepsilon \right) \right]. \end{aligned}$$

Letting $K_1 \left(\varepsilon \right) = - \left(2\delta_2 - \delta_1 \left(\varepsilon \right) \right) > 0$, we have, for any ε,

$$\lim_{n\to\infty} P \left[\sup_{\boldsymbol{\Theta}\setminus N\left(\boldsymbol{\theta}_n^p, \varepsilon \right)} \frac{1}{n} \left[\sum_{t=1}^{n} l_t \left(\boldsymbol{\theta} \right) - \sum_{t=1}^{n} l_t \left(\boldsymbol{\theta}_n^p \right) \right] < -K_1 \left(\varepsilon \right) \right] = 1,$$

which proves the consistency condition given by (10). The proof of the other two concentration conditions (11) and (12) can be done similarly and hence omitted.

Proof of Lemma 3.2: See the technical supplement (Li, et al, 2017).

Proof of Lemma 3.3

From the definition

$$\begin{aligned} & P_D = \int -2 \left[\ln p \left(\mathbf{y} | \boldsymbol{\theta} \right) - \ln p \left(\mathbf{y} | \bar{\boldsymbol{\theta}} \right) \right] p \left(\boldsymbol{\theta} | \mathbf{y} \right) \mathrm{d}\boldsymbol{\theta} \\ = \quad & -2 \int \ln p \left(\mathbf{y} | \boldsymbol{\theta} \right) p \left(\boldsymbol{\theta} | \mathbf{y} \right) \mathrm{d}\boldsymbol{\theta} + 2 \ln p \left(\mathbf{y} | \bar{\boldsymbol{\theta}} \right) \\ = \quad & -2 \sum_{t=1}^{n} \int l_t \left(\boldsymbol{\theta} \right) p \left(\boldsymbol{\theta} | \mathbf{y} \right) \mathrm{d}\boldsymbol{\theta} + 2 \ln p \left(\mathbf{y} | \bar{\boldsymbol{\theta}} \right). \end{aligned}$$

By Lemma 3.2 we have

$$\int l_t(\boldsymbol{\theta})\, p(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta} = l_t\left(\widehat{\boldsymbol{\theta}}_n\right) + \frac{1}{n}B_{t,1} + \frac{1}{n^2}\left(B_{t,2} - B_{t,3}\right) + O_p\left(\frac{1}{n^3}\right),$$

where $B_{t,i}$ is defined in Lemma 3.2 for $i = 1, 2, 3$. Note that

$$\begin{aligned}
\frac{1}{n}\sum_{t=1}^{n} B_{t,1} &= -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] - \frac{1}{n}\sum_{t=1}^{n}\nabla l_t\left(\widehat{\boldsymbol{\theta}}_n\right)'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}} \\
&\quad -\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla l_t\left(\widehat{\boldsymbol{\theta}}_n\right) \\
&= -\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)\right] = -\frac{1}{2}P,
\end{aligned}$$

since $\sum_{t=1}^{n}\nabla l_t\left(\widehat{\boldsymbol{\theta}}_n\right) = 0$. For the same reason

$$\begin{aligned}
&\frac{1}{n}\sum_{t=1}^{n} B_{t,2} \\
= &-\frac{5}{16}\mathbf{tr}\left[A_2\right]\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] + \frac{35}{48}A_1\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right] \\
&-\frac{5}{12}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \\
&+\frac{1}{8}\mathbf{tr}\left[\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\otimes vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)\right]\frac{1}{n}\sum_{t=1}^{n}\nabla^4 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)'\right] \\
&-\frac{5}{4}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}} \\
&+\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\frac{1}{n}\sum_{t=1}^{n}\nabla^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}} \\
&+\frac{3}{4}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{1}{n}\sum_{t=1}^{n}\nabla^2 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\right]\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla^2\widehat{p}}{\widehat{p}}\right] \\
= &-\frac{5}{16}\mathbf{tr}\left[A_2\right]P + \frac{35}{48}A_1P - \frac{5}{12}A_1 + \frac{1}{8}\mathbf{tr}\left[A_2\right] \\
&-\frac{3}{4}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}} + \frac{3}{4}P\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla^2\widehat{p}}{\widehat{p}}\right] \\
= &\frac{2-5P}{16}\mathbf{tr}\left[A_2\right] + \frac{-20+35P}{48}A_1 - \frac{3}{4}C_{21} + \frac{3P}{4}C_{22} + \frac{3P}{4}C_{23},
\end{aligned}$$

where C_{21}, C_{22} and C_{23} are defined in Lemma 3.3. Clearly the first two terms are not related to the prior while the last three terms are related to the prior. We can also show that

$$\frac{1}{n}\sum_{t=1}^{n} B_{t,3}$$

$$
\begin{aligned}
&= B_4 \frac{1}{n}\sum_{t=1}^{n} B_{t,1} \\
&= -\frac{P}{2}\left(-\frac{1}{2}\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla^2\widehat{p}}{\widehat{p}}\right]+\frac{1}{2}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\frac{\nabla\widehat{p}}{\widehat{p}}\right) \\
&\quad -\frac{P}{2}\left(-\frac{5}{24}A_1+\frac{1}{8}\mathbf{tr}\left[A_2\right]\right) \\
&= \frac{P}{4}C_{22}+\frac{P}{4}C_{23}-\frac{P}{4}C_{21}+\frac{P}{2}\left(\frac{5}{24}A_1-\frac{1}{8}\mathbf{tr}\left[A_2\right]\right).
\end{aligned}
$$

Hence,

$$
\begin{aligned}
&\sum_{t=1}^{n}\int l_t\left(\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)\mathrm{d}\boldsymbol{\theta} \\
&= \sum_{t=1}^{n} l_t\left(\widehat{\boldsymbol{\theta}}_n\right)-\frac{P}{2} \\
&\quad +\frac{1}{n}\left(\frac{2-5P}{16}\mathbf{tr}\left[A_2\right]+\frac{-20+35P}{48}A_1-\frac{3}{4}C_{21}+\frac{3P}{4}C_{22}+\frac{3P}{4}C_{23}\right) \\
&\quad +\frac{1}{n}\left[-\frac{P}{4}C_{22}-\frac{P}{4}C_{23}+\frac{P}{4}C_{21}-\frac{P}{2}\left(\frac{5}{24}A_1-\frac{1}{8}\mathbf{tr}\left[A_2\right]\right)\right]+O_p\left(n^{-2}\right) \\
&= \sum_{t=1}^{n} l_t\left(\widehat{\boldsymbol{\theta}}_n\right)-\frac{P}{2} \\
&\quad +\frac{1}{n}\left(\frac{1-2P}{8}\mathbf{tr}\left[A_2\right]+\frac{-10+15P}{24}A_1-\frac{3-P}{4}C_{21}+\frac{P}{2}C_{22}+\frac{P}{2}C_{23}\right)+O_p\left(n^{-2}\right).
\end{aligned}
$$

From the stochastic expansion, we have

$$
\begin{aligned}
\overline{\boldsymbol{\theta}}_n &= \widehat{\boldsymbol{\theta}}_n-\frac{1}{n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\nabla\widehat{\pi} \\
&\quad +\frac{1}{2n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)+O_p\left(\frac{1}{n^2}\right),
\end{aligned}
$$

and

$$
\begin{aligned}
\widehat{\boldsymbol{\theta}}_n-\overline{\boldsymbol{\theta}}_n &= \frac{1}{n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\nabla\widehat{\pi} \\
&\quad -\frac{1}{2n}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n^{(3)}\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)+O_p\left(\frac{1}{n^2}\right). \qquad (30)
\end{aligned}
$$

By the Taylor expansion, we get

$$
\begin{aligned}
&\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}_n\right) \\
&= \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right)+\frac{\partial\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right)}{\partial\theta'}\left(\bar{\boldsymbol{\theta}}_n-\widehat{\boldsymbol{\theta}}_n\right) \\
&\quad +\frac{1}{2}\left(\bar{\boldsymbol{\theta}}_n-\widehat{\boldsymbol{\theta}}_n\right)'\frac{\partial^2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\bar{\boldsymbol{\theta}}_n-\widehat{\boldsymbol{\theta}}_n\right)+O_p\left(\frac{1}{n^2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right) + \frac{1}{2n}\nabla\widehat{\pi}'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\nabla\widehat{\pi} \\
&\quad + \frac{1}{8n}vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right)'\frac{1}{n}\sum_{t=1}^{n}\nabla^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \\
&\quad \times\frac{1}{n}\sum_{t=1}^{n}\nabla^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) \\
&\quad - \frac{1}{2n}\nabla\widehat{\pi}'\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1} \\
&\quad \times\frac{1}{n}\sum_{t=1}^{n}\nabla^3 l_t\left(\widehat{\boldsymbol{\theta}}_n\right)'vec\left(\bar{\mathbf{H}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)^{-1}\right) + O_p(n^{-2}) \\
&= \ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right) - \frac{1}{2n}C_{21} + \frac{1}{2n}C_{23} + \frac{1}{8n}A_1 + O_p(n^{-2}). \qquad (31)
\end{aligned}
$$

Hence, from (30) and (31) we have,

$$
\begin{aligned}
P_D &= -2\sum_{t=1}^{n}\int l_t(\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta} + 2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}_n\right) \\
&= -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right) + P \\
&\quad + \frac{1}{n}\left(-\frac{1-2P}{4}\mathbf{tr}\left[A_2\right] - \frac{-10+15P}{12}A_1 + \frac{3-P}{2}C_{21} - PC_{22} - PC_{23}\right) \\
&\quad + 2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) + O_p\left(n^{-2}\right) \\
&= P + \frac{1}{n}\left(-C_{21} + C_{23} + \frac{1}{4}A_1\right) \\
&\quad + \frac{1}{n}\left(-\frac{1-2P}{4}\mathbf{tr}\left[A_2\right] - \frac{-10+15P}{12}A_1 + \frac{3-P}{2}C_{21} - PC_{22} - PC_{23}\right) + O_p\left(n^{-2}\right) \\
&= P + \frac{1}{n}\left(-\frac{1-2P}{4}\mathbf{tr}\left[A_2\right] - \frac{-13+15P}{12}A_1\right) \\
&\quad + \frac{1}{n}\left(\frac{1-P}{2}C_{21} - PC_{22} + (1-P)\,C_{23}\right) + O_p\left(n^{-2}\right) \\
&= P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(n^{-2}\right), \qquad (32)
\end{aligned}
$$

where C_1 and C_2 are defined in Lemma 3.3. From the definition of DIC, (31) and (32), we have

$$
\begin{aligned}
\mathrm{DIC} &= -4\sum_{t=1}^{n}\int l_t(\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta} + 2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}\right) \\
&= -2\sum_{t=1}^{n}\int l_t(\boldsymbol{\theta})\,p(\boldsymbol{\theta}|\mathbf{y})\,\mathrm{d}\boldsymbol{\theta} + P_D \\
&= -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right) + P + \frac{1}{n}\left(-\frac{1-2P}{4}\mathbf{tr}\left[A_2\right] - \frac{-10+15P}{12}A_1 + \frac{3-P}{2}C_{21} - PC_{22} - PC_{23}\right) \\
&\quad + P + \frac{1}{n}\left(-\frac{1-2P}{4}\mathbf{tr}\left[A_2\right] - \frac{-13+10P}{12}A_1 + \frac{1-P}{2}C_{21} - PC_{22} + (1-P)\,C_{23}\right) + O_p\left(n^{-2}\right)
\end{aligned}
$$

$$
\begin{aligned}
&= -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n\right)+2P \\
&\quad +\frac{1}{n}\left[-\frac{1-2P}{2}\mathbf{tr}\left[A_2\right]-\frac{-23+25P}{12}A_1+(2-P)C_{21}-2PC_{22}+(1-2P)C_{23}\right]+O_p\left(n^{-2}\right) \\
&= \mathrm{AIC}+\frac{1}{n}D_1+\frac{1}{n}D_2+O_p\left(n^{-2}\right),
\end{aligned}
$$

where D_1 and D_2 are defined in Lemma 3.3.

Proof of Theorem 3.1

We write $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right)$ as $\mathbf{H}_n$, $\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)$ as $\mathbf{B}_n$, and let $\mathbf{C}_n=\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}$. Under Assumptions 1-10, we can show that

$$
\overline{\boldsymbol{\theta}}_n(\mathbf{y})=\widehat{\boldsymbol{\theta}}_n(\mathbf{y})+O_p\left(n^{-1}\right) \tag{33}
$$

by (30). Then we have

$$
\begin{aligned}
&\overline{\boldsymbol{\theta}}_n(\mathbf{y})=\boldsymbol{\theta}_n^p+O_p\left(n^{-1/2}\right), \\
&\frac{1}{\sqrt{n}}\mathbf{B}_n^{-1/2}\frac{\partial\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p)}{\partial\boldsymbol{\theta}}\xrightarrow{d}N\left(0,\mathbf{I}_P\right),
\end{aligned} \tag{34}
$$

and

$$
\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\xrightarrow{d}N\left(0,\mathbf{I}_P\right). \tag{35}
$$

Note that

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

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

$$
\begin{aligned}
&\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) \\
= &\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)+\frac{\partial\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) \\
&+\frac{1}{2}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)'\frac{\partial\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)+o_p\left(1\right) \\
= &\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)+\frac{\partial\ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) \\
&+\left(\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)-\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)'\frac{\partial\ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) \\
&+\frac{1}{2}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)'\frac{\partial\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p-\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)+o_p\left(1\right)
\end{aligned}
$$

$$= \ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) + \frac{1}{2}\left(\boldsymbol{\theta}_n^p - \overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)' \frac{\partial \ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\boldsymbol{\theta}_n^p - \overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) + o_p\left(1\right).$$

from (33). Then we have

$$\begin{aligned}
T_2 &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) + 2\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)\right] \\
&= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-\left(\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}_n^p\right)' \frac{\partial \ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}_n^p\right) + o_p\left(1\right)\right] \\
&= E_{\mathbf{y}_{rep}}\left[-\left(\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}_n^p\right)' \frac{\partial \ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}_n^p\right)\right] + o\left(1\right) \\
&= E_{\mathbf{y}}\left[-\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' \frac{\partial \ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n(\mathbf{y}))}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o\left(1\right),
\end{aligned}$$

by Assumption 5-6 and the dominated convergence theorem (Chung, 2001 and DasGupta, 2008). Next, we expand $\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)$ at $\boldsymbol{\theta}_n^p$:

$$\begin{aligned}
\ln p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right) &= \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) + \frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right) \\
&\quad + \frac{1}{2}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' \frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right) + o_p\left(1\right).
\end{aligned}$$

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}}_n(\mathbf{y})\right)) - E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(-2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)) \\
&= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\begin{array}{c} -2\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right) - \\ \left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' \frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right) + o_p\left(1\right) \end{array}\right] \\
&= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] \\
&\quad + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' \frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o\left(1\right) \\
&= -2E_{\mathbf{y}_{rep}}\left(\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\right) E_{\mathbf{y}}\left[\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] \\
&\quad + E_{\mathbf{y}}\left[-\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' E_{\mathbf{y}_{rep}}\left(\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o\left(1\right) \\
&= E_{\mathbf{y}}\left[-\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)' E_{\mathbf{y}}\left(\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_n^p)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y}) - \boldsymbol{\theta}_n^p\right)\right] + o(1),
\end{aligned}$$

since

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

by (34), (35), and the dominated convergence theorem.

Note that

$$\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} = E_{\mathbf{y}}\left(\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right) + o_p\left(1\right),$$

under Assumption 1-10 and by the uniform law of large numbers. Hence, we get

$$\begin{aligned} T_2 &= E_{\mathbf{y}}\left[-\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\frac{\partial \ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n(\mathbf{y})\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right]+o\left(1\right) \\ &= E_{\mathbf{y}}\left[-\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\frac{1}{n}E_{\mathbf{y}}\left(\frac{\partial^2 \ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)+o_p\left(1\right)\right]+o\left(1\right) \\ &= T_3+o(1). \end{aligned}$$

So we only need to analyze T_3. Note that

$$\begin{aligned} T_3 &= E_{\mathbf{y}}\left[-\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'E_{\mathbf{y}}\left(-\frac{1}{n}\frac{\partial^2 \ln p(\mathbf{y}|\boldsymbol{\theta}_n^p)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right]+o\left(1\right) \\ &= E_{\mathbf{y}}\left[\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\left(-\mathbf{H}_n\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right]+o\left(1\right) \\ &= E_{\mathbf{y}}\left[\left(\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right)'\mathbf{C}_n^{1/2}\left(-\mathbf{H}_n\right)\mathbf{C}_n^{1/2}\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\right]+o\left(1\right) \\ &= E_{\mathbf{y}}\left\{\mathbf{tr}\left[\mathbf{H}_n\mathbf{C}_n^{1/2}\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\mathbf{C}_n^{-1/2}\mathbf{C}_n^{1/2}\right]\right\}+o\left(1\right) \\ &= \mathbf{tr}\left\{\left(-\mathbf{H}_n\right)\mathbf{C}_n^{1/2}E_{\mathbf{y}}\left[\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\mathbf{C}_n^{-1/2}\right]\mathbf{C}_n^{1/2}\right\}+o\left(1\right) \\ &= \mathbf{tr}\left\{\left(-\mathbf{H}_n\right)\mathbf{C}_n^{1/2}E_{\mathbf{y}}\left[\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\mathbf{C}_n^{-1/2}\right]\mathbf{C}_n^{1/2}\right\}+o\left(1\right), \end{aligned}$$

and

$$E_{\mathbf{y}}\left[\mathbf{C}_n^{-1/2}\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\overline{\boldsymbol{\theta}}_n(\mathbf{y})-\boldsymbol{\theta}_n^p\right)'\mathbf{C}_n^{-1/2}\right]=\mathbf{I}_P+o\left(1\right).$$

Hence,

$$\begin{aligned} T_3 &= \mathbf{tr}\left(\left(-\mathbf{H}_n\right)\mathbf{C}_n^{1/2}\mathbf{C}_n^{1/2}\right)+o\left(1\right)=\mathbf{tr}\left(\left(-\mathbf{H}_n\right)\mathbf{C}_n\right)+o\left(1\right) \\ &= \mathbf{tr}\left(\left(-\mathbf{H}_n\right)\left(-\mathbf{H}_n\right)^{-1}\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right)+o\left(1\right) \\ &= \mathbf{tr}\left(\left(-\mathbf{H}_n\right)\left(-\mathbf{H}_n\right)^{-1}\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right)+o\left(1\right) \\ &= \mathbf{tr}\left(\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right)+o\left(1\right), \end{aligned}$$

and

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

under the condition that $\mathbf{B}_n+\mathbf{H}_n=o(1)$.

Following Lemma 3.3, we get $P_D=P+o_p(1)$. Finally, we have

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

$$= \quad E_{\mathbf{y}}\left[D\left(\overline{\boldsymbol{\theta}}_n\right) + 2P_D + o_p(1)\right] = E_{\mathbf{y}}\left[\mathrm{DIC} + o_p(1)\right] = E_{\mathbf{y}}\left[\mathrm{DIC}\right] + o(1),$$

by Assumption 10 and the dominated convergence theorem.

Proof of Theorem 4.1

Denote

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

Then we have the following three lemmas under the condition that $\mathbf{y}$ and $\mathbf{y}_{rep}$ are independent. These three lemma are useful to prove Theorem 4.1.

Lemma 5.1 *Under Assumptions 1-7 and 10,* $\widetilde{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_n^p$.

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

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

Let $\delta > 0$ be a number to be set later. Because $\boldsymbol{\Theta}$ is compact, there exists a finite number of spheres of radius δ about $\boldsymbol{\theta}_j$, say $\zeta_\delta\left(\boldsymbol{\theta}_j\right) = \{\boldsymbol{\theta}\in\boldsymbol{\Theta} : \|\boldsymbol{\theta}-\boldsymbol{\theta}_j\| \le \delta\}$, $j = 1, \ldots, K\left(\delta\right)$, which covers $\boldsymbol{\Theta}$ (Gallant and White, 1988). Set $\zeta_j = \zeta_\delta\left(\boldsymbol{\theta}_j\right)$**,** $K = K\left(\delta\right)$**.** It follows that

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

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

$$\begin{aligned} &\left|Q_n\left(\boldsymbol{\theta}\right) - \bar{Q}_n\left(\boldsymbol{\theta}\right)\right| \\ \le\ &\left|Q_n\left(\boldsymbol{\theta}\right) - Q_n\left(\boldsymbol{\theta}_j\right)\right| + \left|Q_n\left(\boldsymbol{\theta}_j\right) - \bar{Q}_n\left(\boldsymbol{\theta}_j\right)\right| + \left|\bar{Q}_n\left(\boldsymbol{\theta}_j\right) - \bar{Q}_n\left(\boldsymbol{\theta}\right)\right| \\ \le\ &\frac{1}{n}\sum_{t=1}^n \left|l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right) - l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_j\right)\right| + \frac{1}{n}\sum_{t=1}^n \left|l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}\right) - l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_j\right)\right| \\ &+ \left|\frac{1}{n}\sum_{t=1}^n \left(l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_j\right) - E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right| + \left|\frac{1}{n}\sum_{t=1}^n \left(l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_j\right) - E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right| \\ &+ \frac{1}{n}\sum_{t=1}^n \left|E\left[l_t\left(\boldsymbol{\theta}\right)\right] - E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right| + \frac{1}{n}\sum_{t=1}^n \left|E\left[l_t\left(\boldsymbol{\theta}\right)\right] - E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right| + \left|\frac{\ln p\left(\boldsymbol{\theta}\right)}{n}\right|, \end{aligned}$$

where $E\left[l_t\left(\boldsymbol{\theta}\right)\right] := E\left[l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right)\right] = E\left[l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}\right)\right]$. By Assumption 4, for all $\boldsymbol{\theta} \in \zeta_j$,

$$\left|l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right) - l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_j\right)\right| \le c_t\left(\mathbf{y}^t\right)\|\boldsymbol{\theta}-\boldsymbol{\theta}_j\| \le \delta c_t\left(\mathbf{y}^t\right).$$

and

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

where $\bar{c}_t=E\left[c_t\left(\mathbf{y}^t\right)\right]=E\left[c_t\left(\mathbf{y}_{rep}^t\right)\right]$. Similarly, we have

$$\left|l_t\left(\mathbf{y}_{rep}^t,\boldsymbol{\theta}\right)-l_t\left(\mathbf{y}_{rep}^t,\boldsymbol{\theta}_j\right)\right|\leq\delta c_t\left(\mathbf{y}_{rep}^t\right).$$

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(\mathbf{y}_t^t\right)+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right|+\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t+\frac{\delta}{n}\sum_{t=1}^{n}c_t\left(\mathbf{y}_{rep}^t\right) \\
&+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}_{rep}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right|+\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t+\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(\mathbf{y}^t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right|+2\frac{\delta}{n}\sum_{t=1}^{n}\bar{c}_t \\
&+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(\mathbf{y}_{rep}^t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}_{rep}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right|+\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(\mathbf{y}^t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\right]\right)\right| \\
&+\delta\left|\frac{1}{n}\sum_{t=1}^{n}\left(c_t\left(\mathbf{y}_{rep}^t\right)-\bar{c}_t\right)\right|+\left|\frac{1}{n}\sum_{t=1}^{n}\left(l_t\left(\mathbf{y}_{rep}^t,\boldsymbol{\theta}\right)-E\left[l_t\left(\boldsymbol{\theta}_j\right)\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 4. It follows that

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

Now choose δ such that $0<\delta<\varepsilon/(4\bar{C})$ (i.e., $\varepsilon-2\delta\bar{C}>\varepsilon/2$). Then

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

Next, choose n_0 so that

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

for all $n \geq n_0$ and all $j = 1, \ldots, K$ by Assumptions 1-10 since K is finite. Hence,

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

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

Lemma 5.2 *Under Assumptions 1-10,* $-(2\mathbf{H}_n)^{-1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \xrightarrow{d} N(0, \mathbf{I}_P)$.

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

$$\begin{aligned}
0 &= \frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_n\right) + \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_n\right)\right] + \frac{1}{n}\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right) \\
&= \frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] + \frac{1}{n}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) + \bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right)\right]\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \\
&\quad + \frac{\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right)}{n}
\end{aligned}$$

where $\widetilde{\boldsymbol{\theta}}_{n1}$ is an intermediate value between $\widetilde{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n^p$. It follows that

$$\begin{aligned}
\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) &= \left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) + \bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right)\right]\right)^{-1} \times \\
&\quad \left(n^{-1/2}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] + n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right)\right).
\end{aligned}$$

Under the assumptions, we have

$$n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right) = o_p(1), \quad -n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) \xrightarrow{p} -\mathbf{H}_n,$$

$$-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) \xrightarrow{p} -\mathbf{H}_n, -n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) \xrightarrow{p} -\mathbf{H}_n,$$

$$[-\mathbf{H}_n]^{1/2} n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) \xrightarrow{d} N(0, \mathbf{I}_P), [-\mathbf{H}_n]^{1/2} n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right) \xrightarrow{d} N(0, \mathbf{I}_P).$$

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

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

■

Lemma 5.3 *Under Assumption 1-10, the asymptotic joint distribution of* $\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)$ *and* $\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)$ *is*

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

Proof. By Lemma 5.2, we have

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) = \left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) + \bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right)\right]\right)^{-1}$$
$$\times\left(n^{-1/2}\sum_{t=1}^{n}\left[\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] + n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right)\right),$$

and

$$\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) = \left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n2}\right)\right)^{-1}\left(n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + n^{-1/2}\ln p\left(\widetilde{\boldsymbol{\theta}}_n\right)\right),$$

where $\widetilde{\boldsymbol{\theta}}_{n2}$ is an intermediate value between $\overleftrightarrow{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n^p$. Hence, we have

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

Then we have

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

■

We are now in the position to prove Theorem 4.1. Under Assumptions 1-10, it can be shown that,

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

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

$$p\left(\mathbf{y}_{rep}|\mathbf{y}\right) = \int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta} = \frac{\int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}}{\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}\right)d\boldsymbol{\theta}}$$

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

where

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

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

Note that

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

The first term is

$$
\begin{aligned}
& -\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right|\right) \\
= \; & -\frac{1}{2}\ln\left|-\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\widetilde{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \\
& +\frac{1}{2}\ln\left|-\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} - \frac{1}{n}\frac{\partial \ln p\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\right| \\
= \; & -\frac{1}{2}\ln\left|-\mathbf{H}_n - \mathbf{H}_n\right| + \frac{1}{2}\ln\left|-\mathbf{H}_n\right| + o_p\left(1\right) \\
= \; & -\frac{1}{2}\ln\left(2^P\left|-\mathbf{H}_n\right|\right) + \frac{1}{2}\ln\left|-\mathbf{H}_n\right| + o_p\left(1\right) = -\frac{1}{2}P\ln 2 + o_p\left(1\right).
\end{aligned} \tag{36}
$$

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

The second term is

$$
\begin{aligned}
& -n\hat{h}_N\left(\widetilde{\boldsymbol{\theta}}_n\right) + n\hat{h}_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right) \\
= \; & \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right) + \ln p\left(\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) - \ln p\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right) \\
= \; & \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + o_p\left(1\right) \\
= \; & \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + E_1 + E_2 + o_p\left(1\right),
\end{aligned} \tag{37}
$$

where

$$
E_1 = \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right),\; E_2 = \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right).
$$

We can further decompose E_1 as

$$E_1 = E_{11} + E_{12},$$

where

$$E_{11} = \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right),\ E_{12} = \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right).$$

For E_{11}, we have

$$\begin{aligned} E_{11} &= \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right) \\ &= \frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + \frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)'\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + o_p(1). \end{aligned}$$

Following Assumption 1-10 and Lemma 5.3, we can similarly prove that

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

where

$$\begin{aligned} &\left[\left(-\mathbf{H}_n\right)^{1/2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right) - \boldsymbol{\theta}_n^p\right)\right]'\left(-2\mathbf{H}_n\right)^{1/2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \\ =\ & \left[\left(-\mathbf{H}_n\right)^{1/2}\left(-n^{-1}\sum_{t=1}^{n}\bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n3}\right)\right)^{-1} n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] \\ & \times\left(-2\mathbf{H}_n\right)^{1/2}\left(-n^{-1}\sum_{t=1}^{n}\left[\bigtriangledown^2 l_t\left(\mathbf{y}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right) + \bigtriangledown^2 l_t\left(\mathbf{y}_{rep}^t, \widetilde{\boldsymbol{\theta}}_{n1}\right)\right]\right)^{-1} \\ & \times\left[n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] \\ =\ & \left[\left(-\mathbf{H}_n\right)^{-1/2} n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right]'\left(-2\mathbf{H}_n\right)^{-1/2} \\ & \times\left[n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right) + n^{-1/2}\sum_{t=1}^{n}\bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right] + o_p(1) \end{aligned}$$

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

with $\widetilde{\boldsymbol{\theta}}_{n3}$ lying between $\overleftrightarrow{\boldsymbol{\theta}}_n(\mathbf{y}_{rep})$ and $\boldsymbol{\theta}_n^p$. Hence,

$$
\left[(-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right]' (-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right) \xrightarrow{d} \chi^2(P). \quad (38)
$$

Thus, we have

$$
E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left(\left[(-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right]' (-2\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right)\right) = 0, \quad (39)
$$

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

$$
\begin{aligned}
& E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[\frac{1}{\sqrt{n}} \frac{\partial \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p)}{\partial \boldsymbol{\theta}'} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right] \\
= & E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left[2^{-1/2}\left[(-\mathbf{H}_n)^{1/2} \sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n(\mathbf{y}_{rep}) - \boldsymbol{\theta}_n^p\right)\right]' (-2\mathbf{H}_n)^{1/2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + o_p(1)\right] \\
= & E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left(2^{-1/2}\left[(-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right]' (-2\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}^t, \boldsymbol{\theta}_n^p\right)\right) + \\
& E_{\mathbf{y}} E_{\mathbf{y}_{rep}}\left(2^{-1}\left[(-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right]' (-\mathbf{H}_n)^{-1/2} n^{-1/2} \sum_{t=1}^{n} \bigtriangledown l_t\left(\mathbf{y}_{rep}^t, \boldsymbol{\theta}_n^p\right)\right) + o(1) \\
= & \frac{1}{2} P + o(1).
\end{aligned} \quad (40)
$$

Moreover,

$$
\begin{aligned}
& \frac{1}{2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)' \frac{1}{n} \frac{\partial^2 \ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p)}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \\
= & -\frac{1}{2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)' (-\mathbf{H}_n) \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + o_p(1) \\
= & -\frac{1}{2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)' (-2\mathbf{H}_n)^{1/2} (-2\mathbf{H}_n)^{-1/2} (-\mathbf{H}_n) (-2\mathbf{H}_n)^{-1/2} (-2\mathbf{H}_n)^{1/2} \\
& \times \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + o_p(1) \\
= & -\frac{1}{4}\left[(-2\mathbf{H}_n)^{1/2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right]' (-2\mathbf{H}_n)^{1/2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) + o_p(1)
\end{aligned}
$$

where

$$
\left[(-2\mathbf{H}_n)^{1/2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right]' (-2\mathbf{H}_n)^{1/2} \sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \xrightarrow{d} \chi^2(P). \quad (41)
$$

From (40) and (41)

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_{11}\right)=\left(\frac{1}{2}-\frac{1}{4}\right)P=\frac{1}{4}P+o\left(1\right).$$

For E_{12}, we have

$$E_{12}=\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)-\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)-\frac{1}{\sqrt{n}}\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)$$
$$-\frac{1}{2}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)'\frac{\partial^2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)+o_p\left(1\right).$$

Since

$$-\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)'\frac{\partial^2\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)\xrightarrow{d}\chi^2\left(P\right), \tag{42}$$

$$E_{\mathbf{y}_{rep}}\left(\frac{1}{\sqrt{n}}\frac{\partial\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial\boldsymbol{\theta}}\right)=0,E_{\mathbf{y}_{rep}}\left(\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)\right)=o\left(1\right) \tag{43}$$

from (42), and (43), we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_{12}\right)=\frac{1}{2}P+o\left(1\right).$$

Then

$$\begin{aligned}E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_1\right) &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right)-\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right) \\ &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_{11}+E_{12}\right)=\frac{3}{4}P+o\left(1\right).\end{aligned} \tag{44}$$

Similarly, we can decompose $E_2=-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)+\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right)$ as

$$E_2=E_{21}+E_{22},$$

where

$$E_{21}=\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right)-\ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right),E_{22}=\ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right).$$

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

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_{21}\right)=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(E_{11}\right)=\frac{1}{4}P+o\left(1\right).$$

For E_{22}, we have

$$\begin{aligned}&\ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) \\ =\ &\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)+\frac{1}{\sqrt{n}}\frac{\partial\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}'}\sqrt{n}\left(\boldsymbol{\theta}_n^p-\overleftrightarrow{\boldsymbol{\theta}}_n\right) \\ &+\frac{1}{2}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)'\frac{1}{n}\frac{\partial^2\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)+o_p\left(1\right),\end{aligned} \tag{45}$$

where

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

$$-\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)'\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) \xrightarrow{d} \chi^2(P).$$

Then we can get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{22}) = -\frac{1}{2}P + o(1),$$

and

$$\begin{aligned} E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_2) &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right) \\ &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{21}+E_{22}) = -\frac{1}{4}P + o(1). \end{aligned} \tag{46}$$

Note that

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

by Lemma 3.3. Mimicking the proof of Theorem 3.1, we get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) = E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right] - P. \tag{47}$$

With (36), (44), (46) and (47), we have

$$\begin{aligned} & E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}\ln p(\mathbf{y}_{rep}|\mathbf{y})\right] \\ = \; & E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\ln\left(\frac{\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right|^{-1/2}\exp\left(-nh_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right)}{\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right|^{-1/2}\exp\left(-nh_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right)}\left(1+O_p\left(\frac{1}{n}\right)\right)+O_p\left(\frac{1}{n^2}\right)\right) \\ = \; & -\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right|\right) \\ & + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + \ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right] \\ & + o(1) \\ = \; & -\frac{P}{2}\ln 2 + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right] + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[E_1+E_2\right] + o(1) \\ = \; & -\frac{P}{2}\ln 2 + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right] + \frac{3}{4}P - \frac{1}{4}P + o(1) \\ = \; & E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + \left(\frac{1}{2}-\frac{\ln 2}{2}\right)P\right) + o(1) \\ = \; & E_{\mathbf{y}}\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) - P + \left(\frac{1}{2}-\frac{\ln 2}{2}\right)P + o(1) \\ = \; & E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) - \frac{1+\ln 2}{2}P\right] + o(1), \qquad (48) \\ = \; & E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}_n\right) - \frac{1+\ln 2}{2}P\right] + o(1). \end{aligned}$$

Therefore, $-2\ln p\left(\mathbf{y}|\bar{\boldsymbol{\theta}}_n\right) + (1+\ln 2)P$ is an unbiased estimator of $E_{\mathbf{y}_{rep}}\left(-2\ln p(\mathbf{y}_{rep}|\mathbf{y})\right)$ 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 Chung, K.L. (2001) A course in probability theory. Academic press.

13 DasGupta, A. (2008). Asymptotic theory of statistics and probability. Springer Science & Business Media.

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

15 Geweke, J. and Keane, M. (2001). Computationally intensive methods for integration in econometrics. *Handbook of econometrics*, **5**, 3463-3568.

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

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

18 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.

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

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

21 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.

22 Li, Y., Yu, J. and T. Zeng (2017) Online Supplement to 'Deviance Information Criterion for Model Selection: Justification and Variation', Singapore Management University.

23 Magnus J R, Neudecker H. (1986). Symmetry, 0-1 matrices and Jacobians: A review . *Econometric Theory*, **2(02)**, 157-190.

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

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

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

27 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.

28 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.

29 Schervish, M. J. (2012). Theory of statistics. *Springer Science & Business Media.*

30 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.

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

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

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

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

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

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

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