Posterior-Based Specification Testing and Model Selection

TAO ZENG *Zhejiang University*

This chapter provides an overview of posterior-based specification testing methods and model selection criteria that have been developed in recent years. For the specification testing methods, the first method is the posterior-based version of IOS_A test. The second method is motivated by the power enhancement technique. For the model selection criteria, we first review the deviance information criterion (DIC). We discuss its asymptotic justification and shed light on the circumstances in which DIC fails to work. One practically relevant circumstance is when there are latent variables that are treated as parameters. Another important circumstance is when the candidate model is misspecified. We then review DIC_L for latent variable models and DIC_M for misspecified models.

11.1 Introduction

For many widely used models in economics and finance, it is not straightforward to obtain the maximum likelihood (ML) estimate (MLE) or construct a nonparametric estimate. Examples include, but are not limited to, latent variable models and structural dynamic choice models; see Imai et al. (2009), Norets (2009). As a result of the difficulties in implementing these frequentist methods, there has been an increasing interest in using Bayesian posterior methods, such as Markov chain Monte Carlo (MCMC) and Sequential Monte Carlo (SMC), to conduct posterior analysis of econometric models. With the rapid growth in computer capability, fitting models of increasing complexity using MCMC or SMC has become more feasible. After the MCMC and SMC output are obtained, two important questions naturally arise. The first question is how to perform the specification test of the model. The second is how to compare alternative models that are not necessarily nested. Specification testing and model selection are of fundamental importance in empirical studies. Therefore, posterior-based answers to these questions are critical in practice.

Li et al. (2018) proposed two new specification tests based on posterior output. The first test (referred to as BIMT) is the posterior-based version of IOS_A of Presnell and Boos (2004) and its asymptotic null distribution is normal. To implement this test, one does not need to specify a model in the alternative hypothesis. The second (referred to as BMT) is motivated by the power enhancement technique of Fan et al. (2015) that is based on the model expansion strategy. It combines a component (called J_1) that tests a null point hypothesis in an expanded model and a power enhancement component (called J_0) obtained from the first test. It has been shown that J_0 converges to zero when the null model is correctly specified and diverges when the null model is misspecified. It has also been shown that J_1 is asymptotically χ^2-distributed, suggesting that the test is asymptotically pivotal, when the null model is correctly specified. The second test has several nice properties. First, its size distortion is small, and hence, the use of bootstrap methods is not necessary. Second, it is straightforward to compute from posterior output, and hence, is applicable to a wide range of models, including latent variable models for which ML and bootstrap methods are difficult to use, and structural dynamic choice models. Third, when the test statistic rejects the specification of a null model and J_1 takes on a large value, BMT suggests the source of misspecification.

In the Bayesian community, there are two important metrics used for model selection. One is the Bayes factor (BF) which compares the posterior model probabilities of candidate models, all conditional on the data. Despite the appeal of its statistical interpretation, the BF suffers from a few serious theoretical and computational difficulties. For example, it is not well-defined under improper priors. It is subject to Jeffreys-Lindley-Bartlett's paradox, that is, it tends to reject the null hypothesis even when the null is correct when a vague prior is used. For many models, the posterior model probabilities, and hence, the BF, are difficult to compute.

The second method is the Deviance information criterion (DIC) of Spiegelhalter et al. (2002). DIC has been interpreted as a Bayesian version of the well-known Akaike information criterion (AIC) of Akaike (1973). Like AIC, DIC is used to select a model that minimizes a plug-in predictive loss. Compared with the BF, DIC can be well defined under improper priors and immune to Jeffreys-Lindley-Bartlett's paradox. More importantly, DIC is easier to calculate from posterior output than the BF, especially in the context of latent variable models. Hence, it has been used in many applications. For example, it has been widely applied to compare alternative specifications in stochastic volatility models (Chan and Grant (2016), and Berg et al. (2004)), and VAR models (Chan and Eisenstat (2018)). However, Spiegelhalter et al. (2014) pointed out that DIC lacks formal theoretical justification.

Li et al. (2020b) provided a frequentist justification for DIC by showing that DIC is an asymptotically unbiased estimator of the expected Kullback-Leibler (KL) divergence between the data generating process (DGP) and a predictive distribution with the posterior mean plugged in. The justification relies on three important conditions. The first condition is that the Bernstein-von

Mises theorem must be valid. The second condition is that the standard ML large sample theory (such as consistency and the asymptotic normality) must be valid. The third condition is that all candidate models are correctly specified, at least asymptotically. These conditions may not hold in practice. Li et al. (2020a) pointed out that the Bernstein-von Mises theorem and the standard ML large sample theory may not hold for the latent variables in latent variable models when DIC is calculated based on the conditional likelihood (i.e., the probability of observed data conditional on the original model parameter and the latent variables). As a result, Li et al. (2020a) proposed a new version of DIC, namely DIC_L, to compare latent variable models. Under a set of regularity conditions, Li et al. (2020a) provided a frequentist justification for DIC_L, similar to that of DIC in Li et al. (2020b). Moreover, Li et al. (2020a) proposed another version of DIC, namely DIC_M, to compare misspecified models. It has been shown that DIC_M can be regarded as a Bayesian version of the TIC of Takeuchi (1976).

The chapter is organized as follows: in Section 2, a review of the posterior-based Specification tests and their asymptotic properties is presented; in Section 3, an overview of the posterior-based model selection criteria, including DIC, DIC_L for latent variable models, and DIC_M for misspecified models is given; Section 4 summarizes the conclusions.

11.2 Posterior-based Specification Tests

It is well-known in the literature that, when the ML method is applied to estimate a candidate model, several specification tests may be used. These include the information matrix test (IMT) of White (1982)), the in-and-out likelihood ratio (IOS) and IOS_A tests of Presnell and Boos (2004). Unfortunately, these tests are not directly applicable to estimates based on posterior output.

Recently, Li et al. (2018) proposed two posterior-based specification tests. One of them can be regarded as a posterior-based version of IOS. The other is constructed by the power enhancement technique of Fan et al. (2015) that can assess the validity of the model specification and identify the source of model misspecification if the null model is rejected.

First, we define some notation. Let $\mathbf{y} = (y_1, \ldots, y_n)'$ be the observed data from a probability measure $\Pr_0$ on the probability space $(\Omega, F, \Pr_0)$. Let model Pr be a collection of candidate models indexed by parameters $\boldsymbol{\theta}$ whose dimension is P. Let $\Pr_{\boldsymbol{\theta}}$ denote Pr indexed by $\boldsymbol{\theta}$. Following White (1987), if there exists $\boldsymbol{\theta}$, such that $\Pr_0 \in \Pr_{\boldsymbol{\theta}}$, we say the model Pr is correctly specified. However, if for all $\boldsymbol{\theta}$, $\Pr_0 \notin \Pr_{\boldsymbol{\theta}}$, we say the model Pr is misspecified. We would like to test whether or not the model in question is correctly specified. Let $g(\mathbf{y})$ be the data generating process (DGP) of $\mathbf{y}$.

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 log-likelihood for the t^{th} observation for any $1 \leq t \leq n$. Often, we simply write $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right)$ as $l_t(\boldsymbol{\theta})$ when there is no ambiguity so that $\ln p(\mathbf{y}|\boldsymbol{\theta}) = \sum_{t=1}^{n} l_t(\boldsymbol{\theta})$. It is important to note that in our definition of the log-likelihood, we ignore the initial condition $\ln p(y_0)$. For weakly dependent data, the impact of the initial condition is asymptotically negligible. We define $l_t^{(j)}(\boldsymbol{\theta})$ to be the j^{th} derivative of $l_t(\boldsymbol{\theta})$ and $l_t^{(j)}(\boldsymbol{\theta}) = l_t(\boldsymbol{\theta})$ when $j = 0$. Additionally, we define

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

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

We impose the following regularity conditions which will be used throughout the Chapter.

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

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

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

Assumption 4: For $j = 0, 1, 2$, for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \boldsymbol{\Theta}$, $\left\| l_t^{(j)}(\boldsymbol{\theta}) - l_t^{(j)}(\boldsymbol{\theta}') \right\| \leq c_t^j(\mathbf{y}^t) \left\| \boldsymbol{\theta} - \boldsymbol{\theta}' \right\|$ in probability, where $c_t^j(\mathbf{y}^t)$ is a positive random variable with $\sup_t \left\| c_t^j(\mathbf{y}^t) \right\|_1 < \infty$ and $\frac{1}{n}\sum_{t=1}^{n}\left(c_t^j(\mathbf{y}^t) - E\left(c_t^j\right.\right.$ 0.

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

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

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

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

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

$$\lim_{n\to\infty} \sup \sup_{\boldsymbol{\Theta}\setminus N(\boldsymbol{\theta}_n^p,\varepsilon)} \frac{1}{n}\sum_{t=1}^{n}\left\{E\left[l_t(\boldsymbol{\theta})\right] - E\left[l_t(\boldsymbol{\theta}_n^p)\right]\right\} < 0, \tag{11.2.1}$$

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 the sequence $\{\mathbf{B}_n(\boldsymbol{\theta}_n^p)\}$ is positive definite, both uniformly in n.

Assumption 9: The prior density $p(\boldsymbol{\theta})$ is thrice continuously differentiable and $0 < p(\boldsymbol{\theta}_n^p) < \infty$ uniformly in n. Moreover, there exists an n^* such that, for any $n > n^*$, the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ is proper and $\int \|\boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\mathbf{y})\, d\boldsymbol{\theta} < \infty$.

Assumption 1 is the compactness condition of the parameter space. Assumption 2 implies weak dependence in y_t. Assumption 3 is the continuity and measurability condition for l_t. 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 dominance condition for l_t. Assumption 6 is the weak dependence condition for l_t, especially for the case where l_t is a measurable function of the distant past or future of a mixing process. Assumptions 4-6 imply that $\frac{1}{n}\sum_{t=1}^{n} E\left[l_t(\boldsymbol{\theta})\right]$ is continuous on $\boldsymbol{\Theta}$ uniformly in n and the likelihood function $\frac{1}{n}\sum_{t=1}^{n} l_t(\boldsymbol{\theta})$ converges to $\frac{1}{n}\sum_{t=1}^{n} E\left[l_t(\boldsymbol{\theta})\right]$ uniformly on $\boldsymbol{\Theta}$. Assumption 7 is the identification condition used in Gallant and White (1988). Assumption 1-7 are sufficient conditions for the consistency of standard ML estimator, and the asymptotic normality can be ensured by adding Assumption 8. These assumptions are well-known primitive conditions for developing the ML theory for dependent and heterogeneous data; see, for example, Gallant and White (1988) and Wooldridge (1994). The first part of Assumption 9 ensures that when the sample size increases, the likelihood information dominates the prior information so that the prior information can be ignored asymptotically, the second part ensures the existence of the second order posterior moment.

11.2.1 Posterior-based IOS test

One of the earliest specification tests is proposed by White (1982) based on the information matrix equivalence. Under the null hypothesis that the model is correctly specified, it can be shown that $\mathbf{H}_n(\boldsymbol{\theta}_n^p)+\mathbf{J}_n(\boldsymbol{\theta}_n^p)=0$. Let $d(y_t,\boldsymbol{\theta})=vech\left[\mathbf{h}_t(\boldsymbol{\theta})+\mathbf{s}_t(\boldsymbol{\theta})\mathbf{s}_t'(\boldsymbol{\theta})\right]$ where $vech$ is the column-wise vectorization with the upper portion excluded. White (1982) proposed the following information matrix test

$$\text{IMT}=nD_n\left(\hat{\boldsymbol{\theta}}\right)U_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)D_n\left(\hat{\boldsymbol{\theta}}\right), \tag{11.2.2}$$

where $\hat{\boldsymbol{\theta}}$ is the MLE of $\boldsymbol{\theta}$, and

$$D_n\left(\hat{\boldsymbol{\theta}}\right)=\frac{1}{n}\sum_{t=1}^{n}d\left(y_t,\hat{\boldsymbol{\theta}}\right),\quad \dot{D}_n=\frac{\partial D_n}{\partial\boldsymbol{\theta}},\quad U_n\left(\hat{\boldsymbol{\theta}}\right)=\frac{1}{n}\sum_{t=1}^{n}\nu_t\left(\hat{\boldsymbol{\theta}}\right)\nu_t\left(\hat{\boldsymbol{\theta}}\right)',$$

$$\nu_t\left(\hat{\boldsymbol{\theta}}\right)=d\left(y_t,\hat{\boldsymbol{\theta}}\right)-\dot{D}_n\left(\hat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)\mathbf{s}_t\left(\hat{\boldsymbol{\theta}}\right).$$

Based on a set of regularity conditions, White (1982) showed that IMT$\xrightarrow{d}\chi^2$ as $n\to\infty$ under the null hypothesis.

Presnell and Boos (2004) proposed an alternative test – the IOS test for models with i.i.d. observations,

$$\text{IOS}=\ln\frac{\prod_{t=1}^{n}p\left(y_t|\hat{\boldsymbol{\theta}}\right)}{\prod_{t=1}^{n}p\left(y_t|\hat{\boldsymbol{\theta}}^{(t)}\right)}=\sum_{t=1}^{n}\left[\ln p\left(y_t|\hat{\boldsymbol{\theta}}\right)-\ln p\left(y_t|\hat{\boldsymbol{\theta}}^{(t)}\right)\right],$$

where $\hat{\boldsymbol{\theta}}^{(t)}$ be the MLE of $\boldsymbol{\theta}$ when the tth observation, y_t, is deleted from the full sample. From the predictive perspective, the single likelihood $p\left(y_t|\hat{\boldsymbol{\theta}}^{(t)}\right)$ can be regarded as the predictive likelihood of y_t by all the other observations. It has been shown that the asymptotic form of IOS is

$$\text{IOS}_A=\mathbf{tr}\left[-\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)\bar{\mathbf{J}}_n\left(\hat{\boldsymbol{\theta}}\right)\right], \tag{11.2.3}$$

and IOS$-$IOS$_A=o_p\left(n^{-1/2}\right)$. Like IMT, IOS$_A$ also compares $\bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)$ with $\bar{\mathbf{J}}_n\left(\hat{\boldsymbol{\theta}}\right)$, but in a ratio form instead of in an additive form. Under the null hypothesis, IOS$_A\xrightarrow{p}P$ and $n^{1/2}\left(\text{IOS}_A-P\right)$ converges to a normal distribution with zero mean and finite variance.

IMT, IOS, and IOS$_A$ all suffer from serious bias distortions if the critical values for testing are based on the asymptotic distributions. It is because that these asymptotic distributions poorly approximate their finite sample counterparts. To reduce the size distortion of these tests, bootstrap methods have been proposed to obtain the critical values. Unfortunately, bootstrap methods are computationally demanding in many cases.

Let $V\left(\bar{\boldsymbol{\theta}}\right)=\int\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)'p\left(\boldsymbol{\theta}|\mathbf{y}\right)\mathrm{d}\boldsymbol{\theta}$, a natural posterior-based informative matrix test statistic can be defined as:

$$\text{BIMT}=\mathbf{tr}\left[nV\left(\bar{\boldsymbol{\theta}}\right)\bar{\mathbf{J}}_n\left(\bar{\boldsymbol{\theta}}\right)\right]=n\int\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)'\bar{\mathbf{J}}_n\left(\bar{\boldsymbol{\theta}}\right)\left(\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)\mathrm{d}\boldsymbol{\theta}, \tag{11.2.4}$$

Based on Assumptions 1-10, Li et al (2018) showed that under the null hypothesis,

$$\text{BIMT}=\text{IOS}_A+O_p\left(n^{-1}\right) \tag{11.2.5}$$

because

$$V\left(\bar{\boldsymbol{\theta}}\right)=-\frac{1}{n}\bar{\mathbf{H}}_n^{-1}(\bar{\boldsymbol{\theta}})+O_p(n^{-2}),$$

see Li et al. (2020a) for detailed proof. From (11.2.5), if the model is correctly specified,

$$\text{BIMT} = P + O_p\left(n^{-1/2}\right) + O_p\left(n^{-1}\right) = P + O_p\left(n^{-1/2}\right),$$

since $n^{1/2}\left(\text{IOS}_A - P\right)$ converges to a normal distribution Assumptions 1-8. Then it is straightforward to show that

$$n^{1/2}\left(\text{BIMT}/P - 1\right) = n^{1/2}\left(\text{IOS}_A/P - 1\right) + o_p\left(1\right),$$

so $n^{1/2}\left(\text{BIMT}/P - 1\right)$ has the same asymptotic distribution as $n^{1/2}\left(\text{IOS}_A/P - 1\right)$. Hence, BIMT may be regarded as the posterior-based version of IOS_A. Unfortunately but not surprisingly, BIMT inherits the size distortion problem of IOS_A and bootstrap methods must be used.

11.2.2 Posterior-based specification test with power enhancement

To solve the size distortion and the computational problem of BIMT, Li et al. (2018) proposed a new posterior-based misspecification test (denoted as BMT) by using the power enhancement technique of Fan et al. (2015).

Power enhancement technique

Fan et al. (2015) considered the hypothesis testing problem of $H_0 : \boldsymbol{\theta} = \mathbf{0}$ where $\boldsymbol{\theta}$ is a high-dimensional vector. The alternative hypothesis H_1 is sparse so that the null hypothesis is violated by only a few components. They showed that traditional tests, such as the Wald test, have low power. To enhance the power, they introduced a power enhancement component which is zero under the null hypothesis with high probability and diverges quickly under sparse alternatives. The new test statistic (call it J) has the form of

$$J = J_0 + J_1,$$

where J_1 is an asymptotically pivotal test statistic, such as Wald test, and J_0 is a power enhancement component. J_0 needs to satisfy three properties: (a) $J_0 \geq 0$ almost surely; (b) under H_0, $\Pr(J_0 = 0|H_0) \to 1$; (c) J_0 diverges in probability under some specific regions of H_1. Clearly, property (a) ensures that J is at least as powerful as J_1; property (b) guarantees that the asymptotic distribution of J under H_0 is determined by J_1 and hence the size of J is asymptotically equivalent to that of J_1; property (c) guarantees that the power of J improves that of J_1.

Posterior-based specification test with power enhancement

Similar to Fan et al. (2015), BMT has two components, J_1 and J_0. To construct J_1, Li et al. (2018) expand $p(\mathbf{y}|\boldsymbol{\theta})$, the original model, to a larger model denoted by $p\left(\mathbf{y}|\boldsymbol{\theta}_L\right)$ where $\boldsymbol{\theta}_L = \left(\boldsymbol{\theta}', \boldsymbol{\theta}_E'\right)'$ with $\boldsymbol{\theta}_E$ being a P_E-dimensional vector. So the expanded model $p\left(\mathbf{y}|\boldsymbol{\theta}_L\right)$ nests the original model $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$, if the specification $p\left(\mathbf{y}|\boldsymbol{\theta}\right)$ is correct, then the true value of $\boldsymbol{\theta}_E$ is zero. Let

$$\begin{aligned}
\mathbf{s}\left(\mathbf{y}, \boldsymbol{\theta}_L\right) &= \frac{\partial \ln p\left(\mathbf{y}|\boldsymbol{\theta}_L\right)}{\partial \boldsymbol{\theta}_L}, \quad C\left(\mathbf{y}, \boldsymbol{\theta}_L\right) = \mathbf{s}\left(\mathbf{y}, \boldsymbol{\theta}_L\right)\mathbf{s}\left(\mathbf{y}, \boldsymbol{\theta}_L\right)', \\
V\left(\bar{\boldsymbol{\theta}}_L\right) &= E\left[\left(\boldsymbol{\theta}_L - \bar{\boldsymbol{\theta}}_L\right)\left(\boldsymbol{\theta}_L - \bar{\boldsymbol{\theta}}_L\right)'|\mathbf{y}\right] = \int \left(\boldsymbol{\theta}_L - \bar{\boldsymbol{\theta}}_L\right)\left(\boldsymbol{\theta}_L - \bar{\boldsymbol{\theta}}_L\right)' p(\boldsymbol{\theta}_L|\mathbf{y})\mathrm{d}\boldsymbol{\theta}_L,
\end{aligned}$$

where $\bar{\boldsymbol{\theta}}_L$ is the posterior mean of $\boldsymbol{\theta}_L$ in the expanded model. Then the J_1 component is constructed as

$$J_1 = \mathbf{tr}\left\{C_E\left(\mathbf{y}, \left(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_E = 0\right)\right) V_E\left(\bar{\boldsymbol{\theta}}_L\right)\right\}, \tag{11.2.6}$$

where $C_E\left(\mathbf{y}, \left(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_E = 0\right)\right)$ is the submatrix of $C\left(\mathbf{y}, \boldsymbol{\theta}_L\right)$ corresponding to $\boldsymbol{\theta}_E$ evaluated at $\left(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_E = 0\right)$ and $V_E\left(\bar{\boldsymbol{\theta}}_L\right)$ is the submatrix of $V\left(\boldsymbol{\theta}_L\right)$ corresponding to $\boldsymbol{\theta}_E$ evaluated at $\bar{\boldsymbol{\theta}}_L$. As shown in Li

et al. (2015), J_1 is a posterior-version of LM test (Breusch and Pagan (1980)) designed to test the point null hypothesis $\boldsymbol{\theta}_E = 0$. They showed that $J_1 \xrightarrow{d} \chi^2(q_E)$ when $\boldsymbol{\theta}_E = 0$. Typically, J_1 has good size property as it is designed to test the point null hypothesis.

If J_1 rejects the hypothesis $\boldsymbol{\theta}_E = 0$, it suggests that the original model $p(\mathbf{y}|\boldsymbol{\theta})$ is misspecified and indicates the source of model misspecification in $p(\mathbf{y}|\boldsymbol{\theta})$. Unfortunately, if J_1 fails to reject the hypothesis $\boldsymbol{\theta}_E = 0$, no conclusion can be drawn about the validity of the original model $p(\mathbf{y}|\boldsymbol{\theta})$. This is because, in practice, there are many different avenues for expanding the model. While J_1 may have sufficient power in some cases, it may have low power in others. This problem is similar to that of the Wald statistic in the context of testing a high-dimensional vector against sparse alternatives, as well explained in Fan et al. (2015). To deal with this problem of low power, they used BIMT to construct J_0, that is,

$$J_0 = \sqrt{n}(\text{BIMT}/P - 1)^2. \tag{11.2.7}$$

Then from (11.2.6) and (11.2.7), the power enhancement posterior-based test for model misspecification is

$$\text{BMT} = J_1 + J_0 = \mathbf{tr}\left\{C_E\left(\mathbf{y}, \left(\bar{\boldsymbol{\theta}}, \boldsymbol{\theta}_E = 0\right)\right) V_E\left(\bar{\boldsymbol{\theta}}_L\right)\right\} + \sqrt{n}(\text{BIMT}/P - 1)^2. \tag{11.2.8}$$

Under some mild regularity conditions, when the model is correctly specified, Li et al. (2018) showed that

$$J_1 \xrightarrow{d} \chi^2(P_E), J_0 = o_p(1), \text{BMT} \xrightarrow{d} \chi^2(P_E).$$

If the model is misspecified with $P^* \neq P$, the order of J_0 takes the form

$$J_0 = \sqrt{n}\left[P^*/P - 1\right]^2 + 2\sqrt{n}\left(P^*/P - 1\right) o_p(1) + O_p(n^{-1/2}) = O_p(\sqrt{n}),$$

where $P^* := \mathbf{tr}\left[-\mathbf{H}\left(\boldsymbol{\theta}_n^p\right)^{-1} \mathbf{J}\left(\boldsymbol{\theta}_n^p\right)\right]$. Then the order of the power of BMT is no less than $O_p(\sqrt{n})$.

BMT has several nice properties. First, compared with IM, IOS, and IOS$_A$, BMT is based on the MCMC output. When the optimization of the likelihood function is difficult but the MCMC draws are available, BMT is easier to compute than IM, IOS, and IOS$_A$. Second, when $\sqrt{n}(\text{BIMT}/q - 1)^2$ does not have the size distortion problem, it is likely that BMT will not suffer from size distortion. As a result, we do not need to use bootstrap method, then avoid intensive computational effort. In addition, by incorporating BIMT into J_0, there is no need to get the asymptotic variance of BIMT which is complicated and difficult to calculate.

11.3 Posterior-based Model Selection Criteria

Model selection is a highly important statistical inference in practice. Many penalty-based information criteria have been proposed to select from candidate models in the literature. A famous example is AIC which requires that MLE is available. The most well-known model selection criterion in the Bayesian framework is DIC of Spiegelhalter et al. (2002), which is constructed based on the posterior distribution of the deviance. It has several desirable features. Firstly, DIC is easy to calculate from the posterior output, such as MCMC output, when the likelihood function has a closed form. Secondly, DIC is applicable to a wide range of statistical models. Third, unlike BFs, DIC is immune to Jeffreys-Lindley-Bartlett's paradox and well defined under improper priors.

In this section, we will review the DIC for Bayesian model selection, especially the frequentist justification of DIC.

11.3.1 DIC for Bayesian model selection

A useful measure of how well the model fits the data, based on both the frequentist and Bayesian approach, is the deviance

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

A small value of $D(\boldsymbol{\theta})$ corresponds to a large value of the log-likelihood, indicating the model fits the data well. Deviance measures the in-sample predictive performance of the model and a more complex model will always yield smaller values than a simple model. Deviance itself is not a good choice for model selection because it does not penalize overfitting, meaning that the most complex model will always be selected. But complex models are not always better since their estimates can be highly variable, for instance, the standard errors of the model parameter can be very large from the frequentist viewpoint and the posterior distributions of parameters may be highly diffuse from the Bayesian viewpoint. For the out-of-sample predictive performance, meaning how well the model predicts future data, simple models often perform better.

The most well-known frequentist approximation to the out-of-sample prediction is the Akaike Information Criterion (AIC):

$$\text{AIC} = -2 \ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right) + 2P,$$

where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the MLE estimator of the parameters, P the number of parameters. The smaller AIC is, the better the model. The first term is deviance evaluated at $\hat{\boldsymbol{\theta}}(\mathbf{y})$, which decreases as the fit of the model improves. $2P$ is called the "penalty" term or the degrees of freedom, which increases as the complexity of the model grows. Thus, in AIC, there is a trade off between model fit and model complexity.

Spiegelhalter et al. (2002) proposed the DIC for Bayesian model comparison. The criterion takes the form of

$$\text{DIC} = D(\bar{\boldsymbol{\theta}}) + 2P_D, \tag{11.3.1}$$

where P_D, used to measure the model complexity and also known as "effective number of parameters", is defined as the difference between the posterior mean of the deviance and the deviance evaluated at the posterior mean of the parameters:

$$P_D = \overline{D(\boldsymbol{\theta})} - D(\bar{\boldsymbol{\theta}}) = -2 \int [\ln p(\mathbf{y}|\boldsymbol{\theta}) - \ln p(\mathbf{y}|\bar{\boldsymbol{\theta}})] p(\boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\boldsymbol{\theta}, \tag{11.3.2}$$

with $\bar{\boldsymbol{\theta}}$ being the posterior mean of $\boldsymbol{\theta}$. Note that $D(\boldsymbol{\theta})$ is a function of $\boldsymbol{\theta}$ thus has a posterior distribution.

Decision-theoretic justification for DIC

Let $g(\mathbf{y})$ be the data generating process of $\mathbf{y}$, $\mathbf{y}_{rep} = (y_{1,rep}, \cdots, y_{n,rep})'$ denote the future replicate data with $\mathbf{y}$, and $\boldsymbol{\theta}_n^p$ 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. The quantity that measures how well a candidate model predicts the replicate data is the KL divergence between $g(\mathbf{y})$ and $p(\mathbf{y}_{rep}|\mathbf{y})$:

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

$$= \int \ln g\left(\mathbf{y}_{rep}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep} - \int \ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right) g\left(\mathbf{y}_{rep}\right) \mathrm{d}\mathbf{y}_{rep}, \tag{11.3.3}$$

where $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ denote a generic predictive distribution. The smaller this KL divergence, the better the candidate model in predicting $g(\mathbf{y}_{rep})$.

If we choose $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ to be the plug-in distribution $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)$ where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ is the MLE estimator of $\boldsymbol{\theta}_n^p$ under the data $\mathbf{y}$, then (11.3.3) can be rewritten as

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

where the expectation $E_{\mathbf{y}}$ and $E_{\mathbf{y}_{rep}}$ are related to $g\left(\mathbf{y}\right)$ and $g\left(\mathbf{y}_{rep}\right)$, respectively. Since $g(\mathbf{y}_{rep})$ is the true DGP and $E_{\mathbf{y}_{rep}}\left(\ln g(\mathbf{y}_{rep})\right)$ is the same across all candidate models, it can be dropped from the above equation.

Assumption 10: $\mathbf{H}_n\left(\boldsymbol{\theta}_n^p\right) + \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) = o\left(1\right)$.

Assumption 10 is a generalization of the definition of "information matrix equality"; see White (1996). It was used as an indicator of no model misspecification. Under Assumptions 1-8 and 10, it is well-known that

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

which means that AIC is an unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ asymptotically, for details, see Burnham and Anderson (2002).

Recently, under Assumption 1-10, Li et al. (2020b) provided a frequentist justification of DIC similar to that of AIC. If the plug-in predictive distribution based on replicate data is $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$ where $\bar{\boldsymbol{\theta}}(\mathbf{y})$ is the posterior mean of $\boldsymbol{\theta}_n^p$ conditional on the data $\mathbf{y}$, consider the following KL divergence

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

Under Assumptions 1-8 and 10, Li et al. (2020b) showed that

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

which means that 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. The decision-theoretic justification to DIC shows that DIC selects a model that asymptotically minimizes the expected KL divergence between the DGP and the plug-in predictive distribution $p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)$ where the expectation is taken with respect to the DGP.

The conditions under which AIC is asymptotically unbiased are: the candidate models provide a good approximation of the true DGP (Assumption 10), the consistency and asymptotic normality of MLE (Assumption 1-8), and the expression for the asymptotic variance of MLE (Assumption 9). For details, see Li et al. (2020b). A key difference between AIC and DIC is that the plug-in predictive distribution is based on different estimators of $\boldsymbol{\theta}_n^p$. In AIC, the ML estimate, $\hat{\boldsymbol{\theta}}(\mathbf{y})$, is used while in DIC, the Bayesian posterior mean, $\bar{\boldsymbol{\theta}}(\mathbf{y})$, is used. That is why DIC is called a Bayesian version of AIC. Under Assumptions 1-9, Li et al. (2020b) proved that

$$\bar{\boldsymbol{\theta}} = \hat{\boldsymbol{\theta}} + O_p(n^{-1}), \tag{11.3.4}$$

$$\bar{\mathbf{H}}_n\left(\bar{\boldsymbol{\theta}}\right) = \bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right) + o_p\left(1\right). \tag{11.3.5}$$

From (11.3.4), the Bayesian posterior mean has the same asymptotic distribution as the MLE and (11.3.5) ensures the validity of the expression for its asymptotic variance. That is why we can obtain the asymptotic unbiasedness of DIC under Assumptions 1-10.

The effect of prior information on P_D

A useful contribution of DIC is that it provides a way to measure the model complexity when the prior information is incorporated, see Brooks (2002). For AIC, the number of degrees of freedom, P, is used to measure the model complexity. In the Bayesian framework, the prior information often imposes additional restrictions on the parameter space, the degrees of freedom may be reduced using a prior, so P_D may not be close to P for a finite n.

Li et al. (2020b) proposed high order approximation for P_D and DIC to see the effect of prior information. For convenience, we let $\bar{\mathbf{H}}_n^{(j)}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^{n} l_t^{(j)}(\boldsymbol{\theta})$ for $j = 3, 4, 5$. Let $\pi(\boldsymbol{\theta}) = \ln p(\boldsymbol{\theta})$, $p^{(j)}(\boldsymbol{\theta})$, $\pi^{(j)}(\boldsymbol{\theta})$ be the jth order derivatives of $p(\boldsymbol{\theta})$, $\pi(\boldsymbol{\theta})$ for $j = 1, 2$. When there is no ambiguity, we write $\hat{\boldsymbol{\theta}}(\mathbf{y})$ as $\hat{\boldsymbol{\theta}}$. Under some regularity conditions, Li et al. (2020b) proved that

$$P_D = P + \frac{1}{n}C_1 + \frac{1}{n}C_2 + O_p\left(n^{-2}\right), \tag{11.3.6}$$

$$\text{DIC} = \text{AIC} + \frac{1}{n}D_1 + \frac{1}{n}D_2 + O_p\left(n^{-2}\right), \tag{11.3.7}$$

where

$$C_1 = \frac{1}{4}\mathbf{tr}\left[A_2\right] - \frac{1}{6}\mathbf{tr}\left[A_3\right], \quad C_2 = -C_{22},$$

$$\begin{aligned} D_1 &= -\frac{1}{4}A_1 + \frac{1}{2}\mathbf{tr}\left[A_2\right] - \frac{1}{3}\mathbf{tr}\left[A_3\right], \\ D_2 &= C_{21} - 2C_{22} - C_{23}, \end{aligned}$$

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

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

$$A_3 = \bar{\mathbf{H}}_n^{(3)}\left(\hat{\boldsymbol{\theta}}\right)' \left(\bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)^{-1} \otimes \bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)^{-1}\right) \bar{\mathbf{H}}_n^{(3)}\left(\hat{\boldsymbol{\theta}}\right) \bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)^{-1},$$

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

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

where vec is the column-wise vectorization. From (11.3.7), the difference between DIC and AIC is $O_p(n^{-1})$, then DIC can be regarded as a Bayesian version of AIC. The effect of the prior information on P_D can be approximated by $C_2 = -\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)^{-1} \pi^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\right]$ where $\bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)$ and $\pi^{(2)}\left(\hat{\boldsymbol{\theta}}\right)$ represent the information from the model and prior, respectively.

The above analysis can be illustrated by the following simple example

$$y_i = \theta + u_i \tag{11.3.8}$$

where $u_i \sim_{i.i.d.} N(\theta, \lambda^{-1} + \tau_i^{-1})$ for $i = 1, ..., n$. The likelihood function is

$$p(\mathbf{y}|\theta) = \prod_{i=1}^{n} \sqrt{\frac{1}{2\pi}\frac{\tau_i\lambda}{\tau_i + \lambda}} \exp\left(-\frac{1}{2}\frac{\tau_i\lambda}{\tau_i + \lambda}(y_i - \theta)^2\right). \tag{11.3.9}$$

Let the prior of θ be $N(\mu_0, \tau_0^{-1})$. Then the posterior, based on the likelihood function (11.3.9) and the prior of θ, is given by

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{\int p(\mathbf{y}|\theta)p(\theta)d\theta} = \sqrt{\frac{\tau^o}{2\pi}} \exp\left(-\frac{\tau^o}{2}(\theta - \mu^o)^2\right).$$

Hence, the posterior mean is

$$\bar{\theta} = E(\theta|\mathbf{y}) = \mu^o = (\tau^o)^{-1}\left(\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} y_i + \tau_0\mu_0\right), \tau^o = \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} + \tau_0,$$

and the posterior variance is $(\tau^o)^{-1}$. Then we have

$$\begin{aligned} \overline{D(\theta)} &= -2\int \ln p(\mathbf{y}|\theta)p(\theta|\mathbf{y})d\theta \\ &= n\ln 2\pi - \sum_{i=1}^{n} \ln\frac{\tau_i\lambda}{\tau_i + \lambda} - \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}\left((y_i - \mu^o)^2 + \frac{1}{\tau^o}\right), \end{aligned}$$

$$D(\bar{\theta}) = -2\ln p(\mathbf{y}|\bar{\theta}) = n\ln 2\pi - \sum_{i=1}^{n} \ln\frac{\tau_i\lambda}{\tau_i + \lambda} - \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}(y_i - \mu^o)^2,$$

$$\begin{aligned} P_D &= \overline{D(\theta)} - D(\bar{\theta}) = \frac{1}{\tau^o}\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} \\ &= \left(\sum_{i=1}^{n} \frac{\tau_i}{\tau_i + \lambda} + \frac{\tau_0}{\lambda}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_i}{\tau_i + \lambda}, \end{aligned}$$

and

$$\begin{aligned} \text{DIC} &= D(\bar{\theta}) + 2P_D \qquad (11.3.10) \\ &= n\ln 2\pi - \sum_{i=1}^{n} \ln\frac{\tau_i\lambda}{\tau_i + \lambda} - \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}(y_i - \mu^o)^2 \\ &\quad +2\left(\sum_{i=1}^{n} \frac{\tau_i}{\tau_i + \lambda} + \frac{\tau_0}{\lambda}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_i}{\tau_i + \lambda}. \end{aligned}$$

The log-likelihood function is

$$\ln p(\mathbf{y}|\theta) = -\frac{n}{2}\ln 2\pi + \frac{1}{2}\sum_{i=1}^{n} \ln\frac{\tau_i\lambda}{\tau_i + \lambda} - \frac{1}{2}\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}(y_i - \theta)^2, \qquad (11.3.11)$$

the MLE estimator of θ is

$$\hat{\theta} = \left(\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}\right)^{-1} \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} y_i. \qquad (11.3.12)$$

From (11.3.11) and (11.3.12), we have

$$\bar{\mathbf{H}}^{(-2)}\left(\hat{\theta}\right) = -n\left(\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda}\right)^{-1}. \qquad (11.3.13)$$

It can be shown that $C_1 = 0$ since the third and fourth derivatives are both zero. Note that the logarithm of the prior density function is

$$\ln p(\mathbf{y}|\theta) = -\frac{n}{2}\ln 2\pi + \frac{1}{2}\ln \tau_0 - \frac{1}{2}\sum_{i=1}^{n} \tau_0 \left(\theta - \mu_0\right)^2 .$$

Then we have

$$\pi^{(2)}\left(\hat{\theta}\right) = -\tau_0. \tag{11.3.14}$$

From (11.3.13) and (11.3.14),

$$C_2 = -\mathbf{tr}\left[\bar{\mathbf{H}}_n\left(\hat{\boldsymbol{\theta}}\right)^{-1}\pi^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\right] = -\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\right)^{-1}\tau_0.$$

Hence we can show that

$$\begin{aligned}
P_D &= \frac{1}{\tau^o}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda} = \left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0\right)^{-1}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda} \\
&= \left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\frac{\tau_0}{n}\right)^{-1}\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda} \\
&= \left(1+\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\right)^{-1}\frac{\tau_0}{n}\right)^{-1} \\
&= 1-\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\right)^{-1}\frac{\tau_0}{n}+O_p\left(\left[\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\right)^{-1}\frac{\tau_0}{n}\right]^2\right) \\
&= P+\frac{1}{n}C_1+\frac{1}{n}C_2+O_p\left(n^{-2}\right)
\end{aligned}$$

with $C_1 = 0$ and the number of parameters $P = 1$.

For the models that $\overline{D(\theta)}$ and P_D do not have a closed form expression, such as the normal model with unknown sampling precision, equation (11.3.6) provides a generalized method for measuring the effect of the prior on P_D. Spiegelhalter et al. (2002) used some specific tricks to derive the relationship between P_D and P for this kind of model, but these tricks are difficult to use for other models. For more details, see Li et al. (2020b).

11.3.2 DIC for Latent Variable Models and Misspecified Models

In this section, we first discuss how to obtain DIC when the model includes latent variables. Second, we introduce a new version of DIC for misspecified models.

DIC with data augmentation

Let $\mathbf{y} = (y_1, \ldots, y_n)$ denote the observed data generated from a probability measure P_0 on the probability space (Ω, F, P_0) and $\mathbf{z} = (\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_n)'$ be the latent variables. The latent variable model is indexed by the some P-dimensional parameter vector, $\boldsymbol{\theta}$. Furthermore, $p(\mathbf{y}|\boldsymbol{\theta})$ is used to denote the observed-data likelihood function, and $p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})$ is denoted as the complete-data likelihood function. The relationship between these two likelihood functions is given by

$$p(\mathbf{y}|\boldsymbol{\theta}) = \int p(\mathbf{y}, \mathbf{z}|\boldsymbol{\theta})d\mathbf{z}. \tag{11.3.15}$$

The integral in (11.3.15) usually does not have an analytical form. For many latent variable models, the dimension of $\mathbf{z}$ is as high as the sample size of the observed data, making it very difficult to accurately approximate the integral numerically. Consequently, the ML method and hence, AIC are difficult to use because doing so requires the calculation of $p(\mathbf{y}|\boldsymbol{\theta})$ for each value of $\boldsymbol{\theta}$ during numerical optimizations.

For the Bayesian analysis based on the observed-data likelihood function $p(\mathbf{y}|\boldsymbol{\theta})$, one would end up with the same problems as in ML and DIC is also difficult to calculate since $\ln p(\mathbf{y}|\boldsymbol{\theta})$ does not have a closed form either. To facilitate the posterior analysis, the data-augmentation strategy of Tanner and Wong (1987) is often used to augment the parameter space to $(\boldsymbol{\theta}, \mathbf{z})$, changing the likelihood function to $p(\mathbf{y}|\boldsymbol{\theta},\mathbf{z})$ which typically has a closed-form expression. Denote the posterior mean of $\mathbf{z}, \boldsymbol{\theta}$ by $\bar{\mathbf{z}}, \bar{\boldsymbol{\theta}}$, obtained from the joint posterior distribution $p(\boldsymbol{\theta}, \mathbf{z}|\mathbf{y})$. Applying DIC developed earlier to the data-augmented MCMC output leads to

$$\text{DIC}^{DA} = D(\bar{\mathbf{z}}, \bar{\boldsymbol{\theta}}) + 2P_D^{DA}, \tag{11.3.16}$$

$$P_D^{DA} = \overline{D(\mathbf{z}, \boldsymbol{\theta})} - D(\bar{\mathbf{z}}, \bar{\boldsymbol{\theta}}) = -2 \int [\ln p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta}) - \ln p(\mathbf{y}|\bar{\mathbf{z}}, \bar{\boldsymbol{\theta}})] p(\mathbf{z}, \boldsymbol{\theta}|\mathbf{y}) \mathrm{d}\mathbf{z} \mathrm{d}\boldsymbol{\theta}, \tag{11.3.17}$$

where $D(\mathbf{z}, \boldsymbol{\theta}) = -2 \ln p(\mathbf{y}|\mathbf{z}, \boldsymbol{\theta})$ which is typically available in closed-form. This way of calculating DIC is the default choice in WinBUGS, following the suggestion of Spiegelhalter et al. (2002). Clearly, the use of data augmentation not only facilitates MCMC sampling, but also makes DIC easier to calculate from the MCMC output. As acknowledged in Spiegelhalter et al. (2014), this default method for calculating DIC from $p(\mathbf{y}|\boldsymbol{\theta}, \mathbf{z})$ for latent variable models is implemented "only to make the technique computationally feasible".

However, from a theoretical viewpoint, DIC^{DA} has a few problems. Firstly, the dimension of the parameter space is much larger, increasing from P to $n + P$ for some models, including stochastic volatility models. It leads to the well-known incidental problem in econometrics where the information about the incidental parameters stops accumulating after a finite number of observations when the dimension of the parameter space grows proportionally to the number of observations, see Neyman and Scott (1948) and Lancaster (2000). In this case, the MLE estimator and posterior mean are both inconsistent and the Bernstein-von Mises theorem also becomes invalid. Then DIC^{DA} may not provide an asymptotically unbiased estimator of the KL divergence up to a constant.

To understand this problem, let us consider the following random effect model from Celeux et al. (2006),

$$y_i = z_i + \varepsilon_i \tag{11.3.18}$$

where $z_i \sim N(\theta, \lambda^{-1})$ and $\varepsilon_i \sim_{i.i.d.} N(0, \tau_i^{-1})$ for $i = 1, ..., n$, z_i and ε_i are mutually independent for all i. For simplicity, let λ and τ_i be known. Different from Celeux et al. (2003) where an improper prior is used for θ, we assume the prior for θ is $N(\mu_0, \tau_0^{-1})$. If we treat z_i as parameters, then the complete-likelihood function is

$$p(\mathbf{y}|\theta, \mathbf{z}) = \prod_{i=1}^{n} p(\mathbf{y}|\theta, z_i) = \prod_{i=1}^{n} \sqrt{\frac{\tau_i}{2\pi}} \exp\left(-\frac{1}{2}\tau_i (y_i - z_i)^2\right), \tag{11.3.19}$$

and the prior of $\theta, \mathbf{z}$ is

$$p(\theta, \mathbf{z}) = p(\mathbf{z}|\theta) p(\theta) = \prod_{i=1}^{n} p(z_i|\theta) p(\theta).$$

The posterior density is

$$p(\theta, \mathbf{z}|\mathbf{y}) = \frac{p(\mathbf{y}|\theta, \mathbf{z}) p(\theta, \mathbf{z})}{p(\mathbf{y})},$$

where

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

Then, the joint posterior density of $\theta, \mathbf{z}$ is

$$\begin{aligned} & p(\theta, \mathbf{z}|\mathbf{y}) \\ \propto & \left[\prod_{i=1}^{n} \sqrt{\frac{\tau_i}{2\pi}} \exp\left(-\frac{\tau_i}{2}(y_i - z_i)^2\right) \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(z_i - \theta)^2\right)\right] \sqrt{\frac{\mu_0}{2\pi}} \exp\left(-\frac{\tau_0}{2}(\theta - \mu_0)^2\right), \end{aligned}$$

and the posterior densities are

$$\begin{aligned} \theta|\mathbf{y} &\sim \mathbf{N}\left(\mu^o, (\tau^o)^{-1}\right), \\ z_i|\theta, \mathbf{y} &\sim \mathbf{N}\left(\frac{y_i\tau_i + \lambda\theta}{\tau_i + \lambda}, \frac{1}{\tau_i + \lambda}\right), \end{aligned}$$

where

$$\mu^o = (\tau^o)^{-1}\left(\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} y_i + \tau_0\mu_0\right), \tau^o = \sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} + \tau_0.$$

Hence, the posterior mean of $\theta, \mathbf{z}$ are

$$\begin{aligned} \bar{\theta} &= \mu^o, \\ \bar{z}_i &= E\left[E\left(z_i|\theta, \mathbf{y}\right)\right] = \frac{y_i\tau_i + \lambda\mu^o}{\tau_i + \lambda} \text{ for } i = 1, 2, ..., n. \end{aligned}$$

Since $\{z_i\}_{i=1}^n$ are treated as parameters, they are incidental in the sense of Neyman and Scott (1948). From (11.3.19), the MLE of z_i is $\hat{z}_i = y_i = z_i + \varepsilon_i$. While it is correctly centered at z_i, it is inconsistent because $\hat{z}_i - z_i$ does not go to zero in probability as n goes to infinity. For the posterior mean $\bar{z}_i$, we have

$$\bar{z}_i = \frac{\tau_i}{\tau_i + \lambda} y_i + \frac{\lambda}{\tau_i + \lambda}\mu^o = y_i - \frac{\lambda}{\tau_i + \lambda}(y_i - \mu^o),$$

and

$$\begin{aligned} \bar{z}_i - z_i &= -\frac{\lambda}{\tau_i + \lambda} z_i + \frac{\tau_i}{\tau_i + \lambda}\varepsilon_i + \frac{\lambda}{\tau_i + \lambda}\mu^o \\ &= -\frac{\lambda}{\tau_i + \lambda}(z_i - \mu^o) + \frac{\tau_i}{\tau_i + \lambda}\varepsilon_i. \end{aligned}$$

Therefore, the posterior mean $\bar{z}_i$ is neither centered at the MLE nor consistent. Clearly, both the standard ML large sample theory and the Bernstein-von Mises theorem fail to hold. These results are not surprising since only one observation (i.e., y_i) contains information about z_i.

To compute P_D^{DA}, we use

$$\begin{aligned} \overline{D(\mathbf{z}, \theta)} &= n\ln 2\pi - \sum_{i=1}^{n} \ln\tau_i + \int \left[\sum_{i=1}^{n} \tau_i \int (y_i - z_i)^2 p\left(z_i|\theta, \mathbf{y}\right) dz_i\right] p\left(\theta|\mathbf{y}\right) d\theta \quad (11.3.20) \\ &= n\ln 2\pi - \sum_{i=1}^{n} \ln\tau_i + \sum_{i=1}^{n} \frac{\tau_i}{\tau_i + \lambda} \\ &\quad + \sum_{i=1}^{n} \frac{\tau_i\lambda^2}{(\tau_i + \lambda)^2} \left(\sum_{i=1}^{n} \frac{\tau_i\lambda}{\tau_i + \lambda} + \tau_0\right)^{-1} + \sum_{i=1}^{n} \frac{\tau_i\lambda^2}{(\tau_i + \lambda)^2}(y_i - \mu^o)^2, \end{aligned}$$

$$D(\bar{\mathbf{z}}, \bar{\theta}) = n\ln 2\pi - \sum_{i=1}^{n} \ln\tau_i + \sum_{i=1}^{n} \tau_i \left(y_i - \frac{y_i\tau_i + \lambda\mu^o}{\tau_i + \lambda}\right)^2 \quad (11.3.21)$$

$$= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(y_i-\mu^o)^2 .$$

Then

$$\begin{aligned} P_D^{DA} &= \overline{D(\mathbf{z},\theta)} - D(\bar{\mathbf{z}},\bar{\theta}) \\ &= \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \left(\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}+\frac{\tau_0}{\lambda}\right)^{-1}\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}\frac{\lambda}{\tau_i+\lambda}. \end{aligned}$$

Hence, DIC^{DA} takes the form

$$\begin{aligned} \text{DIC}^{DA} &= D(\bar{\mathbf{z}},\bar{\theta}) + 2P_D^{DA} \\ &= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(y_i-\mu^o)^2 \\ &\quad +2\left(\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + (\tau^o)^{-1}\sum_{i=1}^{n}\frac{\tau_i\lambda}{(\tau_i+\lambda)^2}\right). \end{aligned} \tag{11.3.22}$$

The KL divergence is

$$2KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right] = C + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right],$$

where

$$\begin{aligned} & E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right] \\ = \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\left[\tau_i E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\left(y_{i,rep} - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o\right)^2\right]\right] \\ = \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + 2\sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\left(E_{\mathbf{y}}\left[(y_i-\mu^o)^2\right]\right). \end{aligned} \tag{11.3.23}$$

It can be shown that DIC^{DA} takes the form

$$\begin{aligned} E_{\mathbf{y}}\left(\text{DIC}^{DA}\right) &= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\left(E_{\mathbf{y}}\left[(y_i-\mu^o)^2\right]\right) \\ &\quad +2\left(\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \tau^{o-1}\sum_{i=1}^{n}\frac{\tau_i\lambda}{(\tau_i+\lambda)^2}\right). \end{aligned} \tag{11.3.24}$$

From (11.3.23) and (11.3.24),

$$\begin{aligned} E_{\mathbf{y}}\left(\text{DIC}^{DA}\right) &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right] \\ &\quad +2\left(\sum_{i=1}^{n}\left(\frac{\tau_i}{\tau_i+\lambda} - \frac{\tau_i+\lambda}{\lambda}\right) + (\tau^o)^{-1}\sum_{i=1}^{n}\frac{\tau_i\lambda}{(\tau_i+\lambda)^2}\right). \end{aligned}$$

Thus, DIC^{DA} is not an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right]$ in this case. For illustration, let $\tau_i = \lambda = 1$ and $\tau_0 = 0$. Then

$$\begin{aligned} \sum_{i=1}^{n}\left(\frac{\tau_i}{\tau_i+\lambda} - \frac{\tau_i+\lambda}{\lambda}\right) &= -1.5n, \\ (\tau^o)^{-1}\sum_{i=1}^{n}\frac{\tau_i\lambda}{(\tau_i+\lambda)^2} &= \frac{1}{2}. \end{aligned}$$

As a result, the difference between $E_{\mathbf{y}}\left(\text{DIC}^{DA}\right)$ and $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}(\mathbf{y}),\bar{\theta}(\mathbf{y})\right)\right]$ has order $O(n)$. See the appendix for a detailed proof.

Second, sometimes a statistical model without latent variables can be represented by another model with latent variables. For example, the model (11.3.8) can be rewritten as (11.3.18) with latent variables since the observed-likelihood function of (11.3.18) can be obtained by integrating $\mathbf{z}$ from the joint density of $\mathbf{y},\mathbf{z}$ conditional on θ, $p(\mathbf{y},\mathbf{z}|\theta)$

$$\begin{aligned} p(\mathbf{y}|\theta) &= \int p(\mathbf{y},\mathbf{z}|\theta)d\mathbf{z} = \prod_{i=1}^{n}\int p(y_i,z_i|\theta)dz_i \\ &= \prod_{i=1}^{n}\sqrt{\frac{1}{2\pi}\frac{\tau_i\lambda}{\tau_i+\lambda}}\exp\left(-\frac{1}{2}\frac{\tau_i\lambda}{\tau_i+\lambda}(y_i-\theta)^2\right) \end{aligned} \tag{11.3.25}$$

which is just the likelihood function given by (11.3.8). Obviously, (11.3.22) is different from (11.3.10) even if the marginal densities, $p(\mathbf{y})$ from these two models are the same. In practice, it is well-known that the Student t distribution can be rewritten as a normal-inverse-gamma distribution where the variance is assumed to follow an inverse-gamma distribution and hence, is treated as a latent variable. For example, the asset pricing model

$$R_t = \boldsymbol{\beta}'\boldsymbol{F}_t + \varepsilon_t, \varepsilon_t \sim t[\mathbf{0},\boldsymbol{\Sigma},\nu] \tag{11.3.26}$$

where R_t is the excess return of portfolio at period t with $N\times 1$ dimension, $\boldsymbol{F}_t$ a $K\times 1$ vector of factor portfolio excess returns, $\boldsymbol{\beta}$ a $N\times K$ vector of scaled covariances, ε_t the random error, $t=1,2,\cdots,n$, $\boldsymbol{\Sigma}$ a diagonal matrix and $\nu=3$, the degree of freedom of t distribution. It can be shown that (11.3.26) can be rewritten as the normal-inverse-gamma distribution form

$$R_t = \boldsymbol{\beta}'\boldsymbol{F}_t + \boldsymbol{\varepsilon}_t, \boldsymbol{\varepsilon}_t \sim N(\mathbf{0},\boldsymbol{\Sigma}/\omega_t), \omega_t \sim \Gamma\left(\frac{\nu}{2},\frac{\nu}{2}\right) \tag{11.3.27}$$

where ω_t is a latent variable. Models given by (11.3.26) and (11.3.27) are identical. If we apply DIC to the t distribution model and DIC^{DA} to the normal-inverse-gamma distribution model, it often leads to different values even under the same priors, one can refer to Li et al. (2020a) for a detailed illustration.

Third, DIC^{DA} is usually sensitive to transformations of latent variables since the dimension of the parameter space is much larger due to data augmentation. Consider the following alternative model to model (11.3.18)

$$y_i = \ln\eta_i + \varepsilon_i \tag{11.3.28}$$

where $\eta_i \sim LN(\theta,\lambda^{-1})$, $\varepsilon_i \sim_{i.i.d.} N(0,\tau_i^{-1})$ for $i=1,...,n$, η_i and ε_i are mutually independent for all i, LN denotes the log-normal distribution. It is clearly that the two models given by (11.3.18) and (11.3.28) are identical as the logarithm of the log-normal distribution is the normal distribution. The complete-likelihood function of (11.3.28) is

$$p(\mathbf{y}|\theta,\eta) = \prod_{i=1}^{n}p(\mathbf{y}|\theta,\eta_i) = \prod_{i=1}^{n}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{1}{2}\tau_i(y_i-\ln\eta_i)^2\right).$$

If the prior of θ is also $N(\mu_0,\tau_0^{-1})$, then the posterior density of θ,η is

$$\begin{aligned} &p(\theta,\eta|\mathbf{y}) \\ \propto &\left[\prod_{i=1}^{n}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i}{2}(y_i-\ln\eta_i)^2\right)\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\exp\left(-\frac{\lambda}{2}(\ln\eta_i-\theta)^2\right)\right]\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right). \end{aligned}$$

Hence, we have

$$\theta|\mathbf{y} \sim N\left(\mu^o,(\tau^o)^{-1}\right),$$

$$\eta_i|\theta,\mathbf{y} \sim LN\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda},\frac{1}{\tau_i+\lambda}\right),$$

where

$$\mu^o=(\tau^o)^{-1}\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}y_i+\tau_0\mu_0\right),\tau^o=\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0.$$

For $i=1,2,...,n$, the posterior means of θ and η_i are

$$\begin{aligned}\bar{\theta} &= E(\theta|\mathbf{y})=\mu^o\\ \bar{\eta}_i &= E(\eta_i|\mathbf{y})=\exp\left(\frac{1}{2(\tau_i+\lambda)}\right)\left[\exp\left(\mu^{*o}+\frac{1}{2}\tau^{*o-1}\right)\right],\end{aligned}$$

where

$$\mu^{*o}=\frac{y_i\tau_i}{\tau_i+\lambda}+\frac{\lambda}{\tau_i+\lambda}\mu^o,\ \tau^{*o-1}=\left(\frac{\lambda}{\tau_i+\lambda}\right)^2(\tau^o)^{-1}.$$

To compute P_D^{DA}, we have

$$\begin{aligned}&\overline{D(\boldsymbol{\eta},\theta)} \qquad (11.3.29)\\ =\ & n\ln 2\pi-\sum_{i=1}^{n}\tau_i+\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}\\ &+\sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0\right)^{-1}+\sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(y_i-\mu^o)^2\\ =\ & \overline{D(\mathbf{z},\theta)}.\end{aligned}$$

$D(\bar{\boldsymbol{\eta}},\bar{\theta})$ can be expressed as

$$D(\bar{\boldsymbol{\eta}},\bar{\theta})=D(\bar{\mathbf{z}},\bar{\theta})+C_d, \qquad (11.3.30)$$

where

$$\begin{aligned}C_d =\ & -\sum_{i=1}^{n}\tau_i\frac{\lambda}{(\tau_i+\lambda)^2}(y_i-\mu^o)-(\tau^o)^{-1}\sum_{i=1}^{n}\tau_i\left(\frac{\lambda}{\tau_i+\lambda}\right)^3(y_i-\mu^o)\\ &+\sum_{i=1}^{n}\frac{\tau_i}{4(\tau_i+\lambda)}+\sum_{i=1}^{n}\frac{\tau_i}{2(\tau_i+\lambda)}\left(\frac{\lambda}{\tau_i+\lambda}\right)^2(\tau^o)^{-1}+\frac{1}{4}\sum_{i=1}^{n}\tau_i\left(\frac{\lambda}{\tau_i+\lambda}\right)^4(\tau^o)^{-2}.\end{aligned}$$

Let $P_{D,z}^{DA}$ and DIC_z^{DA} be P_D^{DA} and DIC^{DA} for model (11.3.18). Let $P_{D,z}^{DA}$ and DIC_z^{DA} be $P_{D,\eta}^{DA}$, DIC_η^{DA} for model (11.5.6). From (11.3.29) and (11.3.30), we have

$$\begin{aligned}P_{D,\eta}^{DA} &= \overline{D(\eta,\theta)}-D(\bar{\boldsymbol{\eta}},\bar{\theta})=\overline{D(\mathbf{z},\theta)}-D(\bar{\mathbf{z}},\bar{\theta})-C_d\\ &= P_{D,z}^{DA}-C_d,\end{aligned}$$

$$\begin{aligned}\text{DIC}_\eta^{DA} &= D(\bar{\boldsymbol{\eta}},\bar{\theta})+2P_{D,\eta}^{DA}=D(\bar{\mathbf{z}},\bar{\theta})+C_d+2P_{D,z}^{DA}-2C_d\\ &= \text{DIC}_z^{DA}-C_d.\end{aligned}$$

For illustration, let $\tau_i=\tau$, $\rho=\frac{\tau}{\tau+\lambda}$, $\tau_0=0$. Then $\mu^o=\bar{y}$, $\tau^o=n\lambda\rho$. It can be shown that

$$C_d=\frac{1}{4}n\rho+\frac{\lambda}{2(1-\rho)^2}+\frac{\tau(1-\rho)^4}{4n\lambda^2\rho^2}.$$

Clearly, the difference in both P_D^{DA} and DIC^{DA} is very large (with the order $O(n)$) between models (11.3.18) and (11.5.6) although they represent identical models. See the appendix for a detailed proof.

DIC_L for latent variable models

For latent variable models, while DIC^{DA} is easier to calculate, it suffers from several theoretical and practical problems. DIC has rigorous theoretical justification, but it is difficult to compute for latent variable models since the observed likelihood function cannot be expressed in closed form. Li et al. (2020a) introduced a new version of DIC (DIC_L) for latent variable models which has a valid justification and is feasible to compute. DIC_L is given by

$$\text{DIC}_L = D(\bar{\boldsymbol{\theta}}) + 2P_L, \tag{11.3.31}$$

where

$$P_L = \mathbf{tr}\left\{\mathbf{I}(\bar{\boldsymbol{\theta}})V(\bar{\boldsymbol{\theta}})\right\}, \tag{11.3.32}$$

and

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

Li et al (2020b) showed that under some regularity conditions, DIC_L is an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$.

Under some regularity conditions, Li et al. (2020a) proved that

$$P_L = P + \frac{1}{n}C_{1,L} + \frac{1}{n}C_{2,L} + O_p\left(\frac{1}{n^2}\right), \tag{11.3.33}$$

$$\text{DIC}_L = \text{AIC} + \frac{1}{n}D_{1,L} + \frac{1}{n}D_{2,L} + O_p\left(\frac{1}{n^2}\right), \tag{11.3.34}$$

where

$$C_{1,L} = \frac{1}{2}C_{11,L} - \frac{1}{2}C_{12,L}, \quad C_{2,L} = -C_{22,L},$$

$$D_{1,L} = C_{11,L} + \frac{5}{4}C_{12,L}, \quad D_{2,L} = C_{21,L} - 2C_{22,L} - C_{23,L},$$

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

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

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

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

From (11.3.33) and (11.3.34), DIC_L can also be regarded as a Bayesian version of AIC and the effect of the prior on P_L is $C_{2,L} = -\mathbf{tr}\left[\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)\pi^{(2)}\left(\hat{\boldsymbol{\theta}}\right)\right]$.

In the context of latent variable models, while DIC^{DA} is trivial to calculate but cannot be justified, DIC is justified but difficult to compute. DIC_L solves this dilemma because it is justified and straightforward to compute. The corresponding deviance is based on the observed-data likelihood function and the latent variables are not treated as parameters.

To illustrate this idea, let us first consider models given by (11.3.18) and (11.3.28). It can be shown that

$$\mathbf{I}(\bar{\theta}) = -\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}, V(\bar{\theta}) = \left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0\right)^{-1},$$

for both models since they share the same observed-data likelihood. Then we have

$$P_{L,z} \quad = \quad P_{L,\eta} = \sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0\right)^{-1}$$

$$= 1 - \frac{1}{n}\left(\frac{1}{n}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}\right)^{-1}\tau_0 + O_p\left(n^{-2}\right).$$

These two values are close to 1 (the number of parameters in the model) if τ_0 goes to zero which means that the prior is vague.

It is important to point out that DIC_L can be computed from MCMC output. While DIC_L does not treat latent variables as parameters, MCMC output may be obtained based on the data augmentation technique without affecting the asymptotic justification of DIC_L.

DIC for misspecified models

Both DIC and DIC_L require that all candidate models be good approximations to DGP (Assumption 10). In many applications, this requirement is too strong. Li et al. (2020a) relaxed this requirement and proposed a new version of DIC (namely DIC_M) to compare misspecified models.

If a candidate model is misspecified, the expected KL divergence between the DGP and $p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)$ can be expressed as

$$\begin{aligned} & E_{\mathbf{y}}\left\{2\times KL\left[g\left(\mathbf{y}_{rep}\right), p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right]\right\} = 2C + E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right)\right] \\ = \; & 2C + E_{\mathbf{y}}\left\{-2\ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}(\mathbf{y})\right) - 2\mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\}\right\} + o(1), \end{aligned} \tag{11.3.35}$$

where $\hat{\boldsymbol{\theta}}(\mathbf{y})$ denotes the MLE of $\boldsymbol{\theta}_n^p$ in the misspecified model and C is a constant across all candidate models. As before, we write $\hat{\boldsymbol{\theta}}(\mathbf{y})$ as $\hat{\boldsymbol{\theta}}$. Based on (11.3.35), Takeuchi information criterion (TIC) is defined as

$$\text{TIC} = -2\ln p\left(\mathbf{y}|\hat{\boldsymbol{\theta}}\right) + 2P_T, \tag{11.3.36}$$

where P_T is a consistent estimator of $-\mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\}$, see Takeuchi (1976). TIC is an asymptotically unbiased estimator of the expected KL divergence minus $2C$ when a candidate model is misspecified. The penalty term P_T which is a consistent estimator of $-\mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\}$ takes the form

$$P_T = -\mathbf{tr}\left\{\bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right)\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)\right\}, \tag{11.3.37}$$

where $\bar{\mathbf{H}}_n^{-1}\left(\hat{\boldsymbol{\theta}}\right)$ is a consistent estimator for $\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)$, and a heteroskedasticity and autocorrelation consistent (HAC) estimator of $\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)$ is defined by

$$\bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right) = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n}\mathbf{s}_t\left(\hat{\boldsymbol{\theta}}\right)\mathbf{s}_\tau\left(\hat{\boldsymbol{\theta}}\right)' k\left(\frac{t-\tau}{\gamma_n}\right),$$

where $k(\cdot)$ is a kernel function and γ_n is the bandwidth (Newey and West (1987)). We require three more assumptions to ensure the consistency and positive semidefiniteness of $\bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right)$ and the consistency of P_T, for more details, see De Jong and Davidson (2000).

Assumption 11: Assume the kernel function $k(\cdot)\in\mathcal{H}$, where

$$\mathcal{H} = \left\{\begin{array}{c} k(\cdot): R\to[-1,1], k(x) = k(-x), \text{for any } x\in R, \\ \int_{-\infty}^{+\infty}|k(x)|\,dx<\infty,\ \int_{-\infty}^{+\infty}\psi(\xi)\,d\xi<\infty, \\ k(\cdot) \text{ is continuous at 0 and at all but a finite number of points in } R \end{array}\right\},$$

where

$$\psi(\xi) = (2\pi)^{-1}\int_{-\infty}^{+\infty}k(x)\,e^{i\xi x}dx.$$

Assumption 12: The bandwidth parameter γ_n is an increasing function of sample size n and $\gamma_n = o\left(n^{1/2}\right)$.

Assumption 13: The expectation of the score function $E\left(\mathbf{s}_t\left(\boldsymbol{\theta}_n^p\right)\right)=0$ for any t.

Assumption 11 and 12 ensure that $\bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right)$ is positive semidefinite with probability 1 (Andrews, 1991), and together with Assumption 10, $P_T = P + o_p(1)$. The Assumptions 1-8 and 11-13 imply that

$$\bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right) - \mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right) \xrightarrow{p} 0$$

which in turn imply that

$$P_T + \mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\} \xrightarrow{p} 0.$$

If the model is estimated by a Bayesian method, Li et al. (2020a) proved that $nV\left(\bar{\boldsymbol{\theta}}\right)$ and $\bar{\boldsymbol{\Omega}}_n\left(\bar{\boldsymbol{\theta}}\right)$ are consistent estimators of $\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)$ and $\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)$, respectively. Hence, DIC is defined as

$$\text{DIC}_M = D\left(\bar{\boldsymbol{\theta}}\right) + 2P_M \text{ with } P_M = \mathbf{tr}\left\{n\bar{\boldsymbol{\Omega}}_n\left(\bar{\boldsymbol{\theta}}\right)V\left(\bar{\boldsymbol{\theta}}\right)\right\}, \tag{11.3.38}$$

where $\bar{\boldsymbol{\theta}}$ is the posterior mean of $\boldsymbol{\theta}$. Li et al. (2020a) proved that DIC_M is an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$ when a candidate model is misspecified. Under Assumptions 1-9 and 11-12, DIC_M and TIC are asymptotically equivalent, that is

$$P_M = P_T + o_p\left(\frac{1}{n}\right), \text{DIC}_M = \text{TIC} + o_p\left(\frac{1}{n}\right), \tag{11.3.39}$$

and

$$\bar{\boldsymbol{\Omega}}_n\left(\bar{\boldsymbol{\theta}}\right) - \bar{\boldsymbol{\Omega}}_n\left(\hat{\boldsymbol{\theta}}\right) \xrightarrow{p} 0.$$

Thus, DIC_M can be regarded as a Bayesian version of TIC.

The conditions to obtain the asymptotic unbiasedness of TIC include that the consistency and asymptotic normality of MLE (Assumption 1-8), and the consistent estimator of $-\mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\}$ based on MLE (Assumption 11-13). From (11.3.4), the Bayesian posterior mean estimator is consistent and asymptotic normal under Assumption 1-9. And (11.3.39) shows that P_M is a consistent estimator of $-\mathbf{tr}\left\{\mathbf{B}_n\left(\boldsymbol{\theta}_n^p\right)\mathbf{H}_n^{-1}\left(\boldsymbol{\theta}_n^p\right)\right\}$ under Assumption 11-13. That is the intuition why DIC_M is an asymptotically unbiased estimator of $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}(\mathbf{y})\right)\right]$.

Since DIC_M applies to both correctly specified and misspecified models while DIC_L applies only to correctly specified models, it may be attempting to use DIC_M rather than DIC_L to select a model. However, DIC_M requires the Fisher information matrix, while is usually easier to compute than the Hessian information matrix required by DIC_L.

11.4 Concluding Remarks

In this chapter, we provided an overview of some approaches developed in recent years for specification testing and model selection. For specification testing, we summarized two posterior-based tests proposed by Li, et al. (2018) and their asymptotic properties; the first method is the posterior-based version of IOS_A test and the second method was motivated by the power enhancement technique. For model selection, we provided an overview of the well-known DIC and its extensions, such as DIC_L for latent variable models and DIC_M for misspecified models. We showed that these approaches not only have good theoretical properties, but also are reasonably simple to compute from posterior output. Hence, with the advance of MCMC and SMC techniques and expanding computing capabilities, these approaches can be applied for a variety of complex models, especially latent variable models. We also illustrated the problem with the commonly used calculation of DIC in practice based on the random effect model.

11.5 Appendix

11.5.1 DIC^{DA} for the random effect model (11.3.18)

The joint posterior density of $\theta, \mathbf{z}$ is

$$
\begin{aligned}
& p(\theta, \mathbf{z}|\mathbf{y}) \qquad (11.5.1)\\
\propto\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i}{2}(y_i-z_i)^2\right)\sqrt{\frac{\lambda}{2\pi}}\exp\left(-\frac{\lambda}{2}(z_i-\theta)^2\right)\right]\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\
\propto\ & \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[(\tau_i+\lambda)z_i^2-2(\tau_i y_i+\lambda\theta)z_i+\tau_i y_i^2+\lambda\theta^2\right]\right\}\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\
=\ & \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}(\tau_i+\lambda)\left[\left(z_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2-\left(\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2+\frac{\tau_i y_i^2+\lambda\theta^2}{\tau_i+\lambda}\right]\right\}\\
& \times\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\
=\ & \exp\left\{-\sum_{i=1}^{n}\frac{\tau_i+\lambda}{2}\left(z_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[-\frac{(\tau_i y_i+\lambda\theta)^2}{\tau_i+\lambda}+\tau_i y_i^2+\lambda\theta^2\right]\right\}\\
& \times\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\
=\ & \exp\left\{\sum_{i=1}^{n}-\frac{\tau_i+\lambda}{2}\left(z_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2\right\}\exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}(y_i-\theta)^2\right\}\\
& \times\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right).
\end{aligned}
$$

Then the density of z_i conditional on $\theta, \mathbf{y}$ is

$$
z_i|\theta,\mathbf{y}\sim\mathbf{N}\left(\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda},\frac{1}{\tau_i+\lambda}\right). \qquad (11.5.2)
$$

The posterior density of θ conditional on $\mathbf{y}$ can be obtained by integrating out $\mathbf{z}$ from (11.5.1)

$$
\begin{aligned}
p(\theta|\mathbf{y}) \ &\propto\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}(y_i-\theta)^2\right\}\exp\left(-\frac{\tau_0}{2}(\theta-\mu_0)^2\right)\\
&=\ \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}(y_i-\theta)^2-\frac{\tau_0}{2}(\theta-\mu_0)^2\right\}\\
&\propto\ \exp\left\{-\frac{1}{2}\left[\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0\right)\theta^2-2\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}y_i+\tau_0\mu_0\right)\theta\right]\right\},
\end{aligned}
$$

that is,

$$
\theta|\mathbf{y}\sim\mathbf{N}\left(\mu^o,(\tau^o)^{-1}\right), \qquad (11.5.3)
$$

where

$$
\mu^o=(\tau^o)^{-1}\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}y_i+\tau_0\mu_0\right),\tau^o=\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0.
$$

From (11.5.2) and (11.5.3), we have

$$
\begin{aligned}
\bar{\theta} &= \mu^o,\\
\bar{z}_i &= E\left(\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}|\mathbf{y}\right)=\frac{\tau_i y_i+\lambda\mu^o}{\tau_i+\lambda}.
\end{aligned}
$$

To compute DIC^{DA}, we have

$$\begin{aligned}
\overline{D(\mathbf{z},\theta)} &= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \int\left[\sum_{i=1}^{n}\tau_i\int(y_i-z_i)^2\,p\left(z_i|\theta,\mathbf{y}\right)dz_i\right]p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i \\
&\quad + \int\left[\sum_{i=1}^{n}\tau_i\int\left(y_i-\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}+\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}-z_i\right)^2 p\left(z_i|\theta,\mathbf{y}\right)dz_i\right]p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i \\
&\quad + \int\left[\sum_{i=1}^{n}\tau_i\left(y_i-\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}\right)^2+\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}\right]p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\tau_i\int\left(y_i-\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}\right)^2 p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\tau_i\int\left(y_i-\frac{y_i\tau_i}{\tau_i+\lambda}-\frac{\lambda}{\tau_i+\lambda}\theta\right)^2 p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\tau_i\int\left(\frac{\lambda}{\tau_i+\lambda}y_i-\frac{\lambda}{\tau_i+\lambda}\theta\right)^2 p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\int(y_i-\theta)^2\,p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\int(y_i-\mu^o+\mu^o-\theta)^2\,p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} \\
&\quad + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(\tau^o)^{-1} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\left(y_i-\mu^o\right)^2,
\end{aligned}$$

$$\begin{aligned}
D(\overline{\mathbf{z}},\bar{\theta}) &= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i-\frac{y_i\tau_i+\lambda\mu^o}{\tau_i+\lambda}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\left(y_i-\mu^o\right)^2,
\end{aligned}$$

$$\begin{aligned}
P_D^{DA} &= \overline{D(\mathbf{z},\theta)} - D(\overline{\mathbf{z}},\bar{\theta}) \\
&= \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(\tau^o)^{-1} \\
&= \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \left(\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}+\frac{\tau_0}{\lambda}\right)^{-1}\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda}\frac{\lambda}{\tau_i+\lambda},
\end{aligned}$$

$$\text{DIC}^{DA} = D(\overline{\mathbf{z}},\bar{\theta}) + 2P_D^{DA} \tag{11.5.4}$$

$$= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(y_i-\mu^o)^2 + 2\left(\sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}\tau^{o-1}\right).$$

Note that

$$\begin{aligned}
& E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}\left(\mathbf{y}\right),\bar{\theta}\left(\mathbf{y}\right)\right)\right] \qquad (11.5.5)\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\left[\tau_i E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left(y_{i,rep} - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o\right)^2\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left[E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}\left(y_{i,rep} - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o\right)^2\right]\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left[E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}\left(\begin{array}{c} y_{i,rep}-\theta \\ +\theta - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o \end{array}\right)^2\right]\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i E_{\mathbf{y}}\left[E_{\mathbf{y}_{rep}}\left[(y_{i,rep}-\theta)^2\right]\right]\\
& + \sum_{i=1}^{n}\tau_i E_{\mathbf{y}}\left[\left(\theta - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o\right)^2\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda}\\
& + \sum_{i=1}^{n}\tau_i E_{\mathbf{y}}\left[\left(\theta - \frac{\tau_i}{\tau_i+\lambda}y_i - \frac{\lambda}{\tau_i+\lambda}\mu^o\right)^2\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda} + \sum_{i=1}^{n}\tau_i E_{\mathbf{y}}\left[\left(\theta - y_i + \frac{\lambda}{\tau_i+\lambda}\left(y_i-\mu^o\right)\right)^2\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda}\\
& + \sum_{i=1}^{n}\tau_i E_{\mathbf{y}}\left[(\theta-y_i)^2\right] + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}E_{\mathbf{y}}\left[(y_i-\mu^o)^2\right]\\
= \; & n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda}\\
& + \sum_{i=1}^{n}\frac{\tau_i+\lambda}{\lambda} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}E_{\mathbf{y}}\left[(y_i-\mu^o)^2\right],
\end{aligned}$$

where we have used the fact that $y_i \sim N\left(\theta, \tau_i^{-1}+\lambda^{-1}\right)$ and $y_{i,rep} \sim N\left(\theta, \tau_i^{-1}+\lambda^{-1}\right)$.

From (11.5.4) and (11.5.5), we have

$$\begin{aligned}
E_{\mathbf{y}}\left(\mathrm{DIC}^{DA}\right) &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[-2\ln p\left(\mathbf{y}_{rep}|\bar{\mathbf{z}}\left(\mathbf{y}\right),\bar{\theta}\left(\mathbf{y}\right)\right)\right] + 2\sum_{i=1}^{n}\left(\frac{\tau_i}{\tau_i+\lambda} - \frac{\tau_i+\lambda}{\lambda}\right)\\
&\quad + 2\sum_{i=1}^{n}\frac{\tau_i\lambda^2}{(\tau_i+\lambda)^2}(\tau^o)^{-1}.
\end{aligned}$$

11.5.2 DIC^{DA} for the random effect model with the log-normal distribution (11.3.28)

The complete-likelihood function is

$$p(\mathbf{y}|\theta,\eta)=\prod_{i=1}^{n}p(\mathbf{y}|\theta,\eta_i)=\prod_{i=1}^{n}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{1}{2}\tau_i\left(y_i-\ln\eta_i\right)^2\right).$$

Since the prior of θ is $N(\mu_0,\tau_0^{-1})$, the posterior density of θ,η is

$$\begin{aligned}
& p(\theta,\eta|\mathbf{y}) \qquad (11.5.6)\\
\propto\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i}{2}\left(y_i-\ln\eta_i\right)^2\right)\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\exp\left(-\frac{\lambda}{2}\left(\ln\eta_i-\theta\right)^2\right)\right]\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i}{2}\left(y_i-\ln\eta_i\right)^2-\frac{\lambda}{2}\left(\ln\eta_i-\theta\right)^2\right)\right]\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{1}{2}\left[\left(\tau_i+\lambda\right)\left(\ln\eta_i\right)^2-2\left(\tau_i y_i+\lambda\theta\right)\ln\eta_i+\tau_i y_i^2+\lambda\theta^2\right]\right)\right]\\
& \times\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i+\lambda}{2}\left[\left(\ln\eta_i\right)^2-2\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\ln\eta_i+\frac{\tau_i y_i^2+\lambda\theta^2}{\tau_i+\lambda}\right]\right)\right]\\
& \times\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i+\lambda}{2}\left[\left(\ln\eta_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2-\left(\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2+\frac{\tau_i y_i^2+\lambda\theta^2}{\tau_i+\lambda}\right]\right)\right]\\
& \times\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{\eta_i^2 2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\exp\left(-\frac{\tau_i+\lambda}{2}\left[\left(\ln\eta_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2+\frac{\tau_i\lambda}{\left(\tau_i+\lambda\right)^2}\left(y_i-\theta\right)^2\right]\right)\right]\\
& \times\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right)\\
=\ & \left[\prod_{i=1}^{n}\sqrt{\frac{\lambda}{2\pi}}\sqrt{\frac{\tau_i}{2\pi}}\sqrt{\frac{1}{\eta_i^2 2\pi}}\exp\left(-\frac{\tau_i+\lambda}{2}\left(\ln\eta_i-\frac{\tau_i y_i+\lambda\theta}{\tau_i+\lambda}\right)^2\right)\right]\\
& \times\prod_{i=1}^{n}\exp\left(-\frac{1}{2}\frac{\tau_i\lambda}{\tau_i+\lambda}\left(y_i-\theta\right)^2\right)\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right).
\end{aligned}$$

Then we have

$$\begin{aligned}
p(\theta|\mathbf{y}) &= \int p(\theta,\eta|\mathbf{y})d\theta \qquad (11.5.7)\\
&\propto \prod_{i=1}^{n}\exp\left(-\frac{1}{2}\frac{\tau_i\lambda}{\tau_i+\lambda}\left(y_i-\theta\right)^2\right)\sqrt{\frac{\mu_0}{2\pi}}\exp\left(-\frac{\tau_0}{2}\left(\theta-\mu_0\right)^2\right).
\end{aligned}$$

From (11.5.6) and (11.5.7), the posterior densities of θ and η_i can be written as

$$\begin{aligned}
\theta|\mathbf{y} &\sim \mathbf{N}\left(\mu^o,(\tau^o)^{-1}\right),\\
\eta_i|\theta,\mathbf{y} &\sim \mathbf{LN}\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda},\frac{1}{\tau_i+\lambda}\right),
\end{aligned}$$

where

$$\mu^o = (\tau^o)^{-1}\left(\sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}y_i+\tau_0\mu_0\right), \tau^o = \sum_{i=1}^{n}\frac{\tau_i\lambda}{\tau_i+\lambda}+\tau_0,$$

and

$$E\left(\eta_i|\theta,\mathbf{y}\right) = \exp\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}+\frac{1}{2\left(\tau_i+\lambda\right)}\right),$$

$$\begin{aligned} Var\left(\eta_i|\theta,\mathbf{y}\right) &= \left[\exp\left(\frac{1}{\tau_i+\lambda}\right)-1\right]\exp\left(2\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}+\frac{1}{(\tau_i+\lambda)}\right) \\ &= \left[\exp\left(\frac{1}{\tau_i+\lambda}\right)-1\right]\exp\left(\frac{1}{(\tau_i+\lambda)}\right)\exp\left(2\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}\right). \end{aligned}$$

The posterior mean of θ is $\bar{\theta} = E(\theta|\mathbf{y}) = \mu^o$. The mean of η_i conditional on $\theta, \mathbf{y}$ is

$$E\left(\eta_i|\theta,\mathbf{y}\right) = \exp\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}+\frac{1}{2\left(\tau_i+\lambda\right)}\right).$$

Then the posterior mean of η_i can be expressed as

$$\bar{\eta}_i = E\left[E\left(\eta_i|\theta,\mathbf{y}\right)\right] = \exp\left(\frac{1}{2\left(\tau_i+\lambda\right)}\right)E\left[\exp\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}\right)|\mathbf{y}\right], \tag{11.5.8}$$

for $i = 1, 2, ..., n$. Note that

$$\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda} = \frac{y_i\tau_i}{\tau_i+\lambda}+\frac{\lambda}{\tau_i+\lambda}\theta.$$

Then

$$\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}|\mathbf{y} \sim \mathbf{N}\left(\mu^{*o}, \tau^{*o-1}\right),$$

where

$$\mu^{*o} = \frac{y_i\tau_i}{\tau_i+\lambda}+\frac{\lambda}{\tau_i+\lambda}\mu^o,\ \tau^{*o-1} = \left(\frac{\lambda}{\tau_i+\lambda}\right)^2(\tau^o)^{-1}.$$

since $\theta|\mathbf{y} \sim \mathbf{N}\left(\mu^o, (\tau^o)^{-1}\right)$. Hence we have

$$\exp\left(\frac{y_i\tau_i+\lambda\theta}{\tau_i+\lambda}\right)|\mathbf{y} \sim \mathbf{LN}\left(\mu^{*o}, \tau^{*o-1}\right).$$

Then, from (11.5.8), we have

$$\bar{\eta}_i = \exp\left(\frac{1}{2\left(\tau_i+\lambda\right)}\right)\left[\exp\left(\mu^{*o}+\frac{1}{2}\tau^{*o-1}\right)\right]. \tag{11.5.9}$$

To compute P_D^{DA}, we have

$$\begin{aligned} &\overline{D(\boldsymbol{\eta},\theta)} \qquad (11.5.10) \\ =\ & n\ln 2\pi-\sum_{i=1}^{n}\ln\tau_i+\int\left[\sum_{i=1}^{n}\tau_i\int\left(y_i-\ln\eta_i\right)^2 p\left(\eta_i|\theta,\mathbf{y}\right)d\eta_i\right]p\left(\theta|\mathbf{y}\right)d\theta \\ =\ & n\ln 2\pi-\sum_{i=1}^{n}\ln\tau_i \\ & +\int\left[\sum_{i=1}^{n}\tau_i\int\left(y_i-E\left(\ln\eta_i|\theta,\mathbf{y}\right)+E\left(\ln\eta_i|\theta,\mathbf{y}\right)-\ln\eta_i\right)^2 p\left(\eta_i|\theta,\mathbf{y}\right)d\eta_i\right]p\left(\theta|\mathbf{y}\right)d\theta \end{aligned}$$

$$
\begin{aligned}
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\int\left(y_i - E\left(\ln\eta_i|\theta,\mathbf{y}\right)\right)^2 p\left(\theta|\mathbf{y}\right)d\theta + \sum_{i=1}^{n}\tau_i\int Var\left(\ln\eta_i|\theta,\mathbf{y}\right)p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\int\left(y_i - \frac{y_i\tau_i+\lambda\mu^o}{\tau_i+\lambda}\right)^2 p\left(\theta|\mathbf{y}\right)d\theta + \sum_{i=1}^{n}\tau_i\int Var\left(\ln\eta_i|\theta,\mathbf{y}\right)p\left(\theta|\mathbf{y}\right)d\theta \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{\left(\tau_i+\lambda\right)^2}\tau^{o-1} + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{\left(\tau_i+\lambda\right)^2}\left(y_i-\mu^o\right)^2 + \sum_{i=1}^{n}\frac{\tau_i}{\tau_i+\lambda},
\end{aligned}
$$

which is the same as $\overline{D(\mathbf{z},\theta)}$. And $D(\bar{\boldsymbol{\eta}},\bar{\theta})$ is

$$
\begin{aligned}
D(\bar{\boldsymbol{\eta}},\bar{\theta}) &= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i-\ln\bar{\eta}_i\right)^2 \qquad (11.5.11)\\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \ln\left(\exp\left(\frac{1}{2\left(\tau_i+\lambda\right)}\right)\left[\exp\left(\mu^{*o}+\frac{1}{2}\tau^{*o-1}\right)\right]\right)\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \ln\left(\exp\left(\frac{1}{2\left(\tau_i+\lambda\right)}\right)\left[\exp\left(\mu^{*o}+\frac{1}{2}\tau^{*o-1}\right)\right]\right)\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \frac{1}{2\left(\tau_i+\lambda\right)} - \mu^{*o} - \frac{1}{2}\tau^{*o-1}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \frac{1}{2\left(\tau_i+\lambda\right)} - \mu^{*o} - \frac{1}{2}\tau^{*o-1}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \frac{1}{2\left(\tau_i+\lambda\right)} - \frac{y_i\tau_i}{\tau_i+\lambda} - \frac{\lambda}{\tau_i+\lambda}\mu^o - \frac{1}{2}\left(\frac{\lambda}{\tau_i+\lambda}\right)^2\left(\tau^o\right)^{-1}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(y_i - \frac{y_i\tau_i}{\tau_i+\lambda} - \frac{\lambda}{\tau_i+\lambda}\mu^o - \frac{1}{2\left(\tau_i+\lambda\right)} - \frac{1}{2}\left(\frac{\lambda}{\tau_i+\lambda}\right)^2\left(\tau^o\right)^{-1}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\tau_i\left(\frac{\lambda}{\tau_i+\lambda}\left(y_i-\mu^o\right) - \frac{1}{2\left(\tau_i+\lambda\right)} - \frac{1}{2}\left(\frac{\lambda}{\tau_i+\lambda}\right)^2\left(\tau^o\right)^{-1}\right)^2 \\
&= n\ln 2\pi - \sum_{i=1}^{n}\ln\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{\left(\tau_i+\lambda\right)^2}\left(y_i-\mu^o\right)^2 + C_d,
\end{aligned}
$$

where

$$
\begin{aligned}
C_d &= -\sum_{i=1}^{n}\tau_i\frac{\lambda}{\left(\tau_i+\lambda\right)^2}\left(y_i-\mu^o\right) - \left(\tau^o\right)^{-1}\sum_{i=1}^{n}\tau_i\left(\frac{\lambda}{\tau_i+\lambda}\right)^3\left(y_i-\mu^o\right) \\
&\quad + \sum_{i=1}^{n}\frac{\tau_i}{4\left(\tau_i+\lambda\right)} + \sum_{i=1}^{n}\frac{\tau_i}{2\left(\tau_i+\lambda\right)}\left(\frac{\lambda}{\tau_i+\lambda}\right)^2\left(\tau^o\right)^{-1} + \frac{1}{4}\sum_{i=1}^{n}\tau_i\left(\frac{\lambda}{\tau_i+\lambda}\right)^4\left(\tau^o\right)^{-2}.
\end{aligned}
$$

Note that

$$
D(\bar{\mathbf{z}},\bar{\theta}) = n\ln 2\pi - \sum_{i=1}^{n}\tau_i + \sum_{i=1}^{n}\frac{\tau_i\lambda^2}{\left(\tau_i+\lambda\right)^2}\left(y_i-\mu^o\right)^2.
$$

Then, by (11.5.11), we have

$$
D(\bar{\boldsymbol{\eta}},\bar{\theta}) = D(\bar{\mathbf{z}},\bar{\theta}) + C_d. \qquad (11.5.12)
$$

From (11.5.10) and (11.5.11), we can get

$$
P_D^{DA} = \overline{D(\boldsymbol{\eta},\theta)} - D(\bar{\boldsymbol{\eta}},\bar{\theta})
$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \frac{\tau_i}{\tau_i+\lambda} + \sum_{i=1}^{n} \frac{\tau_i \lambda^2}{(\tau_i+\lambda)^2} \left(\sum_{i=1}^{n} \frac{\tau_i \lambda}{\tau_i+\lambda} + \tau_0 \right)^{-1} \\
&\quad + \sum_{i=1}^{n} \tau_i \frac{\lambda}{(\tau_i+\lambda)^2} (y_i - \mu^o) + (\tau^o)^{-1} \sum_{i=1}^{n} \tau_i \left(\frac{\lambda}{\tau_i+\lambda} \right)^3 (y_i - \mu^o) - \sum_{i=1}^{n} \frac{\tau_i}{4(\tau_i+\lambda)} \\
&\quad - \sum_{i=1}^{n} \frac{\tau_i}{2(\tau_i+\lambda)} \left(\frac{\lambda}{\tau_i+\lambda} \right)^2 (\tau^o)^{-1} - \frac{1}{4} \sum_{i=1}^{n} \tau_i \left(\frac{\lambda}{\tau_i+\lambda} \right)^4 (\tau^o)^{-2},
\end{aligned}
$$

$$
\text{DIC}^{DA} = D(\bar{\boldsymbol{\eta}}, \bar{\theta}) + 2P_D^{DA}.
$$

Let $\tau_i = \tau$, $\rho = \frac{\tau}{\tau+\lambda}$, $\tau_0 = 0$. Then $\mu^o = \bar{y}$, $\tau^o = n\lambda\rho$, and

$$
\begin{aligned}
\overline{D(\boldsymbol{\eta}, \theta)} &= n \ln 2\pi - n\tau + n\rho \\
&\quad + \lambda n\rho (1-\rho) (n\lambda\rho)^{-1} + \lambda\rho(1-\rho) \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
&= n \ln 2\pi - n\tau + n\rho + (1-\rho) + \lambda\rho(1-\rho) \sum_{i=1}^{n} (y_i - \bar{y})^2,
\end{aligned}
$$

$$
\begin{aligned}
D(\bar{\boldsymbol{\eta}}, \bar{\theta}) &= n \ln 2\pi - n\tau + \frac{\tau\lambda^2}{(\tau+\lambda)^2} \sum_{i=1}^{n} (y_i - \bar{y})^2 + C_d \\
&= n \ln 2\pi - n\tau + \lambda\rho(1-\rho) \sum_{i=1}^{n} (y_i - \bar{y})^2 + C_d,
\end{aligned}
$$

where

$$
\begin{aligned}
C_d &= -\rho(1-\rho) \sum_{i=1}^{n} (y_i - \bar{y}) - (\tau^o)^{-1} \tau (1-\rho)^3 \sum_{i=1}^{n} (y_i - \bar{y}) + 4n\rho + \frac{1}{2} (n\lambda\rho)^{-1} n\rho (1-\rho)^2 \\
&\quad \frac{1}{4} (n\lambda\rho)^{-2} n\tau (1-\rho)^4 \\
&= \frac{1}{4} n\rho + \frac{\lambda}{2(1-\rho)^2} + \frac{\tau(1-\rho)^4}{4n\lambda^2\rho^2},
\end{aligned}
$$

since $\sum_{i=1}^{n} (y_i - \bar{y}) = 0$. Hence, we have

$$
\begin{aligned}
P_{D,\eta}^{DA} &= \overline{D(\boldsymbol{\eta}, \theta)} - D(\bar{\boldsymbol{\eta}}, \bar{\theta}) = \overline{D(\mathbf{z}, \theta)} - D(\bar{\mathbf{z}}, \bar{\theta}) - C_d \\
&= P_{D,z}^{DA} - C_d,
\end{aligned}
$$

$$
\begin{aligned}
\text{DIC}_{\eta}^{DA} &= D(\bar{\boldsymbol{\eta}}, \bar{\theta}) + 2P_D^{DA} = 2\overline{D(\boldsymbol{\eta}, \theta)} - D(\bar{\boldsymbol{\eta}}, \bar{\theta}) \\
&= 2\overline{D(\mathbf{z}, \theta)} - D(\bar{\mathbf{z}}, \bar{\theta}) - C_d \\
&= \text{DIC}_{z}^{DA} - C_d.
\end{aligned}
$$

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Second international symposium on information theory*, Volume 1, pp. 267–281. Springer Verlag.

Andrews, D. W. (1987). Consistency in nonlinear econometric models: A generic uniform law of large numbers. *Econometrica: Journal of the Econometric Society 55*(6), 1465–1471.

Berg, A., R. Meyer, and J. Yu (2004). Deviance information criterion for comparing stochastic volatility models. *Journal of Business and Economic Statistics 22*(1), 107–120.

Breusch, T. S. and A. R. Pagan (1980). The lagrange multiplier test and its applications to model specification in econometrics. *The Review of Economic Studies 47*(1), 239–253.

Burnham, K. and D. Anderson (2002). *Model selection and multi-model inference: a practical information-theoretic approach*. Springer.

Celeux, G., F. Forbes, C. Robert, and D. Titterington (2006). Deviance information criteria for missing data models. *Bayesian Analysis 1*(4), 651–674.

Chan, J. C. and E. Eisenstat (2018). Bayesian model comparison for time-varying parameter vars with stochastic volatility. *Journal of applied econometrics 33*(4), 509–532.

Chan, J. C. and A. L. Grant (2016). Modeling energy price dynamics: Garch versus stochastic volatility. *Energy Economics 54*, 182–189.

De Jong, R. M. and J. Davidson (2000). Consistency of kernel estimators of heteroscedastic and autocorrelated covariance matrices. *Econometrica 68*(2), 407–423.

Fan, J., Y. Liao, and J. Yao (2015). Power enhancement in high-dimensional cross-sectional tests. *Econometrica 83*(4), 1497–1541.

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

Imai, S., N. Jain, and A. Ching (2009). Bayesian estimation of dynamic discrete choice models. *Econometrica 77*(6), 1865–1899.

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

Li, Y., X. B. Liu, and J. Yu (2015). A bayesian chi-squared test for hypothesis testing. *Journal of Econometrics 189*(1), 54–69.

Li, Y., J. Yu, and T. Zeng (2018). Specification tests based on MCMC output. *Journal of Econometrics 207*(1), 237–260.

Li, Y., J. Yu, and T. Zeng (2020a). Deviance information criterion for latent variable models and misspecified models. *Journal of Econometrics 216*(2), 450–493.

Li, Y., J. Yu, and T. Zeng (2020b). Deviation information criterion: Justification and variation. *Working Paper*.

Newey, W. K. and K. D. West (1987). Hypothesis testing with efficient method of moments estimation. *International Economic Review 28*(3), 777–787.

Neyman, J. and E. L. Scott (1948). Consistent estimates based on partially consistent observations. *Econometrica: Journal of the Econometric Society 16*(1), 1–32.

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

Presnell, B. and D. D. Boos (2004). The IOS test for model misspecification. *Journal of the American Statistical Association 99*(465), 216–227.

Spiegelhalter, D., N. Best, B. Carlin, and A. Van Der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society Series B 64*(4), 583–639.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. Linde (2014). The deviance information criterion: 12 years on. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 76*(3), 485–493.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion of model fitting. *Suri-Kagaku (Mathematic Sciences) 153*(2), 12–18.(in Japanese).

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica 50*(1), 1–25.

White, H. (1987). *Specification testing in dynamic models*, Volume 1. Cambridge University Press New York, NY, USA.

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

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