Robust Bayesian Model Selection*

Yong Li
Renmin University of China

Jun Yu
Singapore Management University

September 30, 2013

Abstract

This paper extends the robust Bayesian inference in misspecified models of Müller (2013, Econometrica) to Bayesian model selection of a set of misspecified models. It is shown that when a model is misspecified, under the Kullback-Leibler loss function, the risk associated with Müller's posterior is less (weakly) than that with the original posterior distribution asymptotically. Based on this new result, two new information criteria are proposed for model selection under model misspecification. Sufficient conditions are provided for the risk associated with Müller's posterior to be strictly smaller.

JEL classification: C11, C12, G12
Keywords: Model selection; Model misspecification; Artificial posterior distribution, Sandwich-covariance matrix; Markov chain Monte Carlo.

Essentially, all models are wrong, but some are useful. (George Box)

1 Introduction

Economic theory often makes strong predictions on certain aspects of economic behavior while at the same time is silent on other aspects. One of the best known cases is that economists are often agnostic about the form of the distribution, especially when distributions are not normally distributed. As a result, robust statistical inference of economic models has received a great deal of attention from econometricians and empirical economists. In frequentist's paradigm, seminar methodological contributions include Huber (1967), Hansen (1982), White (1982), Gourerioux, et al. (1984a, 1984b).

For a long time, robust Bayesian analysis has focused on investigating the sensitivity of posterior distributions to prior distributions, leaving aside the issue of adequacy of the

*Yong Li, Hanqing Advanced Institute of Economics and Finance, Renmin University of China, Beijing, 1000872, P.R. China. Jun Yu, Sim Kee Boon Institute for Financial Economics, School of Economics and Lee Kong Chian School of Business. Yu would like to acknowledge the financial support from Singapore Ministry of Education Academic Research Fund Tier 2 under the grant number MOE2011-T2-2-096. Email for Jun Yu: yujun@smu.edu.sg. URL: http://www.mysmu.edu/faculty/yujun/.

model specification. More recently, the robustness of posteriors was checked in the context of a set of competing models; see for example, Pericchi and Pérez (1994). On the other hand, an extensive literature performs nonparametric Bayesian analysis with the aid of the Dirichlet process (Ferguson, 1973). The applications of this sort of nonparametric Bayesian treatments have generated some fruitful outcomes in economic and financial applications in recent years; see, for example, Jensen and Maheu (2010 and 2013).

None of the above mentioned Bayesian approaches really tells the impact of model specification on the quality of standard Bayesian inferential techniques, as in the way where White (1982) investigated the impact of model specification on maximum likelihood (ML), one of the most important and widely used frequentist's inferential techniques. White showed that there is a discrepancy between the ML theory of the correctly specified model and that of a misspecified model. In the correctly specified model, the ML estimator (MLE) is consistent towards the true value and follows a normal distribution asymptotically whose covariance is the inverse of Fisher information matrix. In a misspecified model, on the other hand, while the asymptotic distribution of MLE is also normal, it centers on the pseudo true value that minimizes the Kullback-Leibler loss between the two models and has a "sandwich" covariance matrix.

To conduct the Bayesian inference, however, the posterior distribution converges to a normal distribution with MLE as its mean and the inverse of Fisher information matrix as its covariance, whether the model is correctly specified or not. Consequently, the Bayesian inference based on this posterior distribution is not robust with respect to model misspecification.

To the best of our knowledge, the first systematic study of the impact of model specification on the quality of the standard Bayesian inferential technique is in Müller (2013) where the author advocates using an artificial posterior distribution, in particular, a normal distribution with MLE as its mean but with a sandwich estimate as its covariance. He then showed that when the model is misspecified, Bayesian inference relying on the new posterior achieves a lower asymptotic frequentist risk than the posterior distribution corresponding to the misspecified model. This result points out an important observation that the traditional Bayesian inferential technique is suboptimal when the model is misspecified.

The present paper reinforces this observation by extending Müller's result to the Bayesian model selection problem. It is shown that when the model is misspecified, under the Kullback-Leibler loss function, the risk of Müller's posterior distribution is lower (weakly) than that of the original posterior distribution, as the sample size goes to infinite. Based on this new result, we then propose two new information criteria for comparing a set of misspecified models. We show that the model selected by the information criteria

that ignores model misspecification can be different from the model selected by our new information criteria. Given that all models are wrong and some are useful, as George Box put it in the header cited above, the information criteria that take model misspecification into account should be very useful in practice. In addition, the proposed information criteria are easy to compute, facilitating implementation in real applications.

This paper is organized as follows. Section 2 introduces the setup and reviews the results of Müller (2013). Section 3 extends the results of Müller to model selection. Section 4 introduces the new information criteria based on the results obtained in Section 3. Section 5 concludes. Appendix collects the proof of the theoretical results of the paper.

2 Bayesian Estimation under Model Misspecification

Let us first fix some notations of this paper. For any $i < j$, we let $\mathbf{y}^{i:j} = (\mathbf{y}_i, \mathbf{y}_{i+1} \cdots, \mathbf{y}_j)$. If $i = 1$, simply write $\mathbf{y}^j = \mathbf{y}^{1:j}$. Let data $\mathbf{y}^n = (\mathbf{y}_1, \mathbf{y}_2, \cdots, \mathbf{y}_n)$ are from the data generating process $p(\mathbf{y}|D)$, where D represents the true model. When there is no confusion, we simply write $\mathbf{y} = \mathbf{y}^n$.

Consider a parametric model, M, denoted by $p(\mathbf{y}|M, \theta)$, where θ is a set of parameters in the model. When there is no confusion, we simply write $p(\mathbf{y}|M, \theta)$ as $p(\mathbf{y}|\theta)$. Denote $\theta_0 \in \Theta \subset R^p$ the pseudo true value that minimizes the Kullback-Leibler (KL) loss between the data generating process and the parametric model,

$$\theta_0 = \arg\min_{\theta} \int \log \frac{p(\mathbf{y}|D)}{p(\mathbf{y}|\theta)} p(\mathbf{y}|D) d\mathbf{y},$$

that is,

$$\int \frac{\partial \log p(\mathbf{y}|\theta)}{\partial \theta} p(\mathbf{y}|D) d\mathbf{y}|_{\theta=\theta_0} = 0.$$

Let $\hat{\theta}$ denote the pseudo ML estimator of θ that maximizes the log-likelihood function of the parametric model,

$$\hat{\theta} = \arg\max_{\theta} \log p(\mathbf{y}|\theta),$$

that is,

$$\frac{\partial \log p(\mathbf{y}|\theta)}{\partial \theta}|_{\theta=\hat{\theta}} = 0.$$

Let $l_t(\theta) = \partial \log p(\mathbf{y}^t|\theta)/\partial\theta$, $s_t(\theta) = l_t(\theta) - l_{t-1}(\theta)$, $h_t(\theta) = \partial s_t(\theta)/\partial\theta$, $\mathbf{J}(\theta) = E_D\left(s_t(\theta)s_t(\theta)'\right)$, $\mathbf{I}(\theta) = -E_D\left(h_t(\theta)\right)$, $\hat{\mathbf{J}}(\theta) = \frac{1}{n}\sum_{t=1}^n s_t(\theta)s_t(\theta)'$, $\hat{\mathbf{I}}(\theta) = -\frac{1}{n}\sum_{t=1}^n h_t(\theta)$. If the model is correctly specified, the pseudo true value becomes the true value. In this case, according to the standard asymptotic ML theory, we have

$$\hat{\theta} \overset{a}{\sim} N\left(\theta_0, \hat{\mathbf{I}}^{-1}(\hat{\theta})/n\right). \tag{1}$$

White (1982) established the misspecified ML theory by showing that

$$\hat{\theta} \overset{a}{\sim} N\left(\theta_0, \hat{\mathbf{I}}^{-1}(\hat{\theta})\hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})/n\right), \tag{2}$$

where the asymptotic variance takes a sandwich form. White's result is more general than that in (1) because when the model is correctly specified, we have the equivalence of the Hessian $\mathbf{J}(\theta_0)$ and the outer product of the score $\mathbf{I}(\theta_0)$.

To conduct Bayesian inference about θ, let $p(\theta)$ be the prior distribution of θ. Given the likelihood function $p(\mathbf{y}|\theta)$, the posterior distribution is:

$$p(\theta|\mathbf{y}) = \frac{p(\mathbf{y}|\theta)p(\theta)}{p(\mathbf{y})} \propto p(\theta)p(\mathbf{y}|\theta), \tag{3}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\theta)p(\theta)d\theta$ is the marginal likelihood. Unlike the case of ML, there is no discrepancy in the Bayesian asymptotic theory between the true model and the misspecified model. In both cases, the large sample Bayesian theory is given by

$$\theta|\mathbf{y} \overset{a}{\sim} N\left(\hat{\theta}, \hat{\mathbf{I}}^{-1}(\hat{\theta})/n\right).$$

The arise of the identical large sample Bayesian theory in the context of misspecification is due to the fact that the sampling distribution of $\hat{\theta}$ and the model likelihood function are different in the case of misspecification. This suggests a better inference about θ may be possible. To do so, Müller (2013) constructed an artificial posterior distribution based on the sandwich covariance matrix in (2). He then showed that the Bayesian inference based on this new posterior leads to a smaller frequentist risk than that on the original posterior under a general class of loss functions. In particular, the new sandwich posterior distribution is given by:

$$p^a(\theta|\mathbf{y}) \propto p(\theta)N\left(\hat{\theta}, \hat{\mathbf{I}}^{-1}(\hat{\theta})\hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})/n\right).$$

Let $d(\mathbf{y})$ denote a Bayes decision (which amounts to the choice of an estimator of θ in this context) and $\mathcal{L}(\theta, d(\mathbf{y}))$ denote a loss function associated with the decision. The Bayes risk under the two different posterior distributions is given by

$$R(d) = \int\int \mathcal{L}(\theta, d(\mathbf{y}))p(\theta|\mathbf{y})d\theta p(\mathbf{y})d\mathbf{y} = E_Y\left(\int \mathcal{L}(\theta, d(\mathbf{y}))p(\theta|\mathbf{y})d\theta\right),$$
$$R^a(d) = \int\int \mathcal{L}(\theta, d(\mathbf{y}))p^a(\theta|\mathbf{y})d\theta p^a(\mathbf{y})d\mathbf{y} = E_{Y^a}\left(\int \mathcal{L}(\theta, d(\mathbf{y}))p^a(\theta|\mathbf{y})d\theta\right),$$

where $p^a(\mathbf{y}) = \int p^a(\theta|\mathbf{y})p(\theta)d\theta$ are the marginal likelihood associated to Müller's posterior distribution. The optimal decision under different Bayes risk can be expressed as:

$$d^*(\mathbf{y}) = \arg\min_d \int \mathcal{L}(\theta, d(\mathbf{y}))p(\theta|\mathbf{y})d\theta, \text{ and } d^{a*}(\mathbf{y}) = \arg\min_d \int \mathcal{L}(\theta, d(\mathbf{y}))p^a(\theta|\mathbf{y})d\theta.$$

Müller (2013) showed that under a general class of loss functions, $d^{a*}(\mathbf{y})$ is superior to $d^*(\mathbf{y})$ in the sense that $r(d^*(\mathbf{y}),\theta) \geq r(d^{a*}(\mathbf{y}),\theta)$ for each $\theta \in \Theta$, where $r(d(\mathbf{y}),\theta)$ is the frequentist risk defined by

$$r(d(\mathbf{y}),\theta) = \int \mathcal{L}(\theta, d(\mathbf{y}))p(\mathbf{y}|D)d\mathbf{y}.$$

This result is important as it not only proves that the standard Bayesian estimator under model misspecification is not optimal, but also offers an improved Bayesian estimator. In our view, it seriously and rigorously addresses a well-known concern, for the first time in the literature, that in robust Bayes studies too little attention has been paid to the robustness with respect to the sampling model, and consequently, it opens the door for robust Bayesian inference under model misspecification.

3 Risk of Alternative Predictive Distributions under Misspecification

Let us consider the problem of how to assess the usefulness of a candidate model. A commonly used method to assess the usefulness of a model is to examine its out-of-the-sample performance. For a candidate model M that is misspecified, corresponding to the two different posterior distributions, i.e., $p(\theta|\mathbf{y})$ (the true posterior distribution associated with model M) and $p^a(\theta|\mathbf{y})$ (Müller's posterior distribution), two different predictive distributions can be constructed. Given some future observations $\mathbf{y}_f$, we define $p(\mathbf{y}_f|\mathbf{y})$ (the regular predictive distribution) and $p^a(\mathbf{y}_f|\mathbf{y})$ (Müller's predictive distribution) by

$$p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\theta,\mathbf{y})p(\theta|\mathbf{y})d\theta, \tag{4}$$

$$p^a(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\theta,\mathbf{y})p^a(\theta|\mathbf{y})d\theta. \tag{5}$$

A natural question to ask is which of these two predictive distributions we should use. In this Section, a comparison is made from the decision-theoretical viewpoint.

To fix the idea, let $\mathbf{y}^{n+1:2n} = (\mathbf{y}_{n+1}, \cdots, \mathbf{y}_{2n})$ denote some future data coming from the true data generating process $p(\mathbf{y}|D)$. For $t = 1, \cdots, n$, let $p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta)$ denote the probability density of $\mathbf{y}_{n+t}$ conditional on $\mathbf{y}^{n+t-1}$. When the misspecification is ignored, the conditional probability distribution of $\mathbf{y}^{n+1:2n}$ is:

$$p(\mathbf{y}^{n+1:2n}|\mathbf{y}) = \prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta)p(\theta|\mathbf{y})d\theta\right]. \tag{6}$$

Based on the posterior distribution of Müller (2013), the conditional probability distribution of $\mathbf{y}^{n+1:2n}$ is:

$$p^a(\mathbf{y}^{n+1:2n}|\mathbf{y}) = \prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta)p^a(\theta|\mathbf{y})d\theta\right]. \tag{7}$$

Both conditional distributions measure the predictive power of the model. When the true data generating process and, hence, the corresponding predictive distribution are known, we can measure the loss of information by examining the difference between each of the conditional distributions and the predictive distribution implied by the true data generating process.

One of the most widely used loss functions is the Kullback-Leibler divergence. For any two distributions $f(x)$ and $g(x)$, KL divergence is defined as:

$$\int \log\left(\frac{f(x)}{g(x)}\right) f(x)dx.$$

In our context, if we choose to use $p(\mathbf{y}^{n+1:2n}|\mathbf{y})$ to measure the predictive power of the model, the loss is

$$\int \log \frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)}{p(\mathbf{y}^{n+1:2n}|\mathbf{y})} p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)d\mathbf{y}^{n+1:2n}.$$

If we choose to use $p^a(\mathbf{y}^{n+1:2n}|\mathbf{y})$ to measure the predictive power of the model, the loss is

$$\int \log \frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)}{p^a(\mathbf{y}^{n+1:2n}|\mathbf{y})} p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)d\mathbf{y}^{n+1:2n}.$$

The risks based on these two loss functions can be expressed as, respectively,

$$\begin{aligned} r_n &= \int \left[\int \log \frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)}{p(\mathbf{y}^{n+1:2n}|\mathbf{y})} p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)d\mathbf{y}^{n+1:2n} \right] p(\mathbf{y}|D)d\mathbf{y} \\ &= \int \log p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)p(\mathbf{y}^{2n}|D)d\mathbf{y}^{2n} - \int \log p(\mathbf{y}^{n+1:2n}|\mathbf{y})p(\mathbf{y}^{2n}|D)d\mathbf{y}^{2n} \\ &= r_{1n} - r_{2n}, \end{aligned}$$

and

$$\begin{aligned} r_n^a &= \int \left[\int \log \frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)}{p^a(\mathbf{y}^{n+1:2n}|\mathbf{y})} p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)d\mathbf{y}^{n+1:2n} \right] p(\mathbf{y}|D)d\mathbf{y} \\ &= r_{1n} - r_{2n}^a. \end{aligned}$$

Obviously, r_{1n} in r_n and r_n^a does not depend on the candidate model.

Defining the risk to be the expectation of the KL loss is a well-known practice in the literature. To name just a few examples, see Akaike (1973), Takeuchi (1976), Laud and Ibrahim (1995), Gelfand and Ghosh (1998), and Spiegelhalter et al. (2002). There is an important difference between our risk function and the risk function used in the literature for the purpose of model selection, however. Our risk function is introduced to measure the expected losses in predicted data sets while the risk function used in much of the literature is the expected losses on replicate data sets.

Our aim is to compare these two risks, r_n and r_n^a. To do so, following Müller (2013), we first impose the following regularity conditions.

Assumption 1: The prior density $p(\theta)$ is continuous and positive at $\theta = \theta_0$. The likelihood dominates the prior information.

Assumption 2: θ_0 is an interior point of θ and $\{l_t\}_{t=1}^n$ is twice continuously differentiable in a neighborhood Θ_0 of θ_0.

Assumption 3: $\sup_{t<n} n^{-1/2}||s_t(\theta_0)|| \xrightarrow{p} 0$, $n^{-1}\sum_{t=1}^n s_t(\theta_0)s_t(\theta_0)' \xrightarrow{p} \mathbf{J}(\theta_0)$, where $\mathbf{J}(\theta_0) \in R^p$ almost surely, and $n^{-1/2}\sum_{t=1}^n s_t(\theta_0) \xrightarrow{d} \mathbf{J}(\theta_0)^{1/2}Z$ with $Z \sim N(0, I_p)$ independent of $\mathbf{J}(\theta_0)$.

Assumption 4: Let $L_n(\theta) = \log p(\theta|\mathbf{y})$. For all $\epsilon > 0$, there exists $K(\epsilon) > 0$ so that, as $n \to \infty$,

$$P_{n,\theta_0}\left(\sup_{||\theta-\theta_0||\geq\epsilon} n^{-1}\left(L_n(\theta) - L_n(\theta_0)\right) < -K(\epsilon)\right) \longrightarrow 1.$$

Assumption 5: $\frac{1}{n}\sum_{t=1}^n ||h_t(\theta_0)|| = O_p(1)$ and $\sup_{t\leq n}||h_t(\theta_0)|| = o_p(n)$. For any sequence $k_n \to 0$,

$$\sup_{||\theta_t-\theta_0||<k_n} n^{-1}\sum_{i=1}^n ||h_t(\theta_t) - h_t(\theta_0)|| \xrightarrow{p} 0,$$

and $n^{-1}\sum_{t=1}^n h_t(\theta_0) \xrightarrow{p} -\mathbf{I}(\theta_0)$ where $\mathbf{I}(\theta_0) \in R^P$ almost surely and $\mathbf{I}(\theta_0)$ is independent of Z. Furthermore, it is assumed that

$$\sup_{||\theta-\theta_0||<k_n}\left[\sup_{t\leq n}||h_t(\theta)||\right] = o_p(n).$$

Assumption 6: Assumptions 1-5 hold for all $\mathbf{y}^{n+t}$, $t = 1, 2, \cdots, n$.

Assumption 7: Assume that the data generating process is strictly stationary and that the following condition holds:

$$\int \frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{1:n+t-1}, \theta)}{\partial\theta} p(\mathbf{y}^{n+1:2n}|\mathbf{y}, D)|_{\theta=\theta_0} = 0.$$

Remark 3.1 *Under Assumptions 1-5, we have*

$$\begin{aligned}\bar{\theta} &= \hat{\theta} + o_p(n^{-1/2}),\\ V(\hat{\theta}) &= -L_n^{-(2)}(\hat{\theta}) + o_p(n^{-1}) = O_p(n^{-1}).\end{aligned}$$

where $\bar{\theta} := E[\theta|\mathbf{y}]$ *and* $V(\hat{\theta}) := E\left[(\theta-\hat{\theta})(\theta-\hat{\theta})'|\mathbf{y}\right]$, $L_n^{(2)}(\theta) := \frac{\partial^2 \log p(\theta|\mathbf{y})}{\partial\theta\partial\theta'}$. *This Bayesian large sample theory has been developed in the literature based on different sets of regularity conditions, see Ghosh and Ramamoorthi (2003), Li, et al. (2013).*

Lemma 3.1 *Let $\tilde{\theta}_{n+t}$ be a consistent estimator of θ_0 constructed from $\mathbf{y}^{1:n+t}$. It can be shown that, under Assumptions 1-5,*

$$\sup_{t\leq n}\left[\frac{1}{n+t}||h_{n+t}(\tilde{\theta}_{n+t})||\right]=o_p(1),$$
$$\sup_{t\leq n}\left[\frac{1}{n+t}||s_{n+t}(\tilde{\theta}_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})||\right]=o_p(1),$$
$$\frac{1}{n}\sum_{t=1}^{n}h_{n+t}(\tilde{\theta}_{n+t})=\frac{1}{n}\sum_{t=1}^{n}h_t(\theta_0)+o_p(1)=-\mathbf{I}(\theta_0)+o_p(1),$$
$$\frac{1}{n}\sum_{t=1}^{n}s_{n+t}(\tilde{\theta}_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})=\frac{1}{n}\sum_{t=1}^{n}s_t(\theta_0)s'_t(\theta_0)+o_p(1)=\mathbf{J}(\theta_0)+o_p(1),$$
$$\frac{1}{n}\sum_{t=1}^{n}||h_{n+t}(\tilde{\theta}_{n+t})||=O_p(1),\frac{1}{n}\sum_{t=1}^{n}||s_{n+t}(\tilde{\theta}_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})||=O_p(1).$$

Lemma 3.2 *Under Assumptions 1-5, we have*

$$\bar{\theta}=\theta_0+O_p(n^{-1/2}),$$
$$V(\theta_0)=V(\bar{\theta})+nV(\bar{\theta})\hat{\mathbf{J}}(\bar{\theta})V(\bar{\theta})+o_p(n^{-1}).$$

Theorem 3.1 *Under Assumptions 1-7, it can be shown that*

$$\lim_{n\to+\infty}r_n^a\leq\lim_{n\to+\infty}r_n.$$

Remark 3.2 *Theorem 4.1 shows that when the model is misspecified, under the KL loss function, the risk of Müller's posterior distribution is less (weakly) than that of the original posterior distribution, as the sample size goes to infinite. This result extends the Müller (2013)'s results to assess the predictive power of a misspecified model.*

4 Bayesian Model Selection under Misspecification

Now suppose there are q candidate models that are all misspecified and we have to select a model. Denote these candidate models by M_j, $j=1,2,\cdots,q$, and let d_j denotes the statistical decision to select model M_j. The model selection problem is to select the optimal model. As argued in the last section, we do so by minimizing the risk of the statistical decision. Following Akaike (1973) and Takeuchi (1976), we assume that parameter θ is only estimated from the sample $\mathbf{y}$. Unlike Akaike (1973) and Takeuchi (1976), we do not plug the MLE into the KL divergence.

If the misspecification is ignored, then the risk associated with decision d_j is

$$r_n(d_j)=\int\left[\int\log\frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)}{p(\mathbf{y}^{n+1:2n}|\mathbf{y},d_j)}p(\mathbf{y}^{n+1:2n}|\mathbf{y},D)d\mathbf{y}^{n+1:2n}\right]p(\mathbf{y}|D)d\mathbf{y}=r_{1n}-r_{2n}(d_j),$$

where

$$p(\mathbf{y}^{n+1:2n}|\mathbf{y}, d_j) = \prod_{t=1}^{n} \left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta) p(\theta|\mathbf{y}, d_j) d\theta \right]. \tag{8}$$

If the misspecification is taken into account and Müller's posterior distribution is used, then the risk associated with decision d_j is

$$r_n^a(d_j) = \int \left[\int \log \frac{p(\mathbf{y}^{n+1:2n}|\mathbf{y}, D)}{p^a(\mathbf{y}^{n+1:2n}|\mathbf{y}, d_j)} p(\mathbf{y}^{n+1:2n}|\mathbf{y}, D) d\mathbf{y}^{n+1:2n} \right] p(\mathbf{y}|D) d\mathbf{y} = r_{1n} - r_{2n}^a(d_j),$$

where

$$p^a(\mathbf{y}^{n+1:2n}|\mathbf{y}, d_j) = \prod_{t=1}^{n} \left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta) p^a(\theta|\mathbf{y}, d_j) d\theta \right]. \tag{9}$$

When the misspecification is ignored, the optimal model is

$$j^* := \arg\min_j r_n(d_j) = \arg\max_j r_{2n}(d_j).$$

When the misspecification is taken into account, the optimal model is

$$j^{a*} := \arg\min_j r_n^a(d_j) = \arg\max_j r_{2n}^a(d_j).$$

Traditionally, model selection has been performed using information criteria that measure the relative quality of a statistical model, for a given set of data. Well-known criteria include Akaike information criterion (AIC) of Akaike (1973), Takeuchi information criterion (TIC) of Takeuchi (1976), Bayesian information criterion (BIC) of Schwarz (1978), posterior information criterion (PIC) of Phillips (1996), and deviance information criterion (DIC) of Spiegelhalter et al. (2002), to name just a few. AIC, TIC and DIC are constructed by estimating the KL divergence and hence share some similarities to our method. We now introduce two new information criteria which can be used to estimate $-2r_{2n}(d_j)$ and $-2r_{2n}^a(d_j)$. Like other information criteria, they can be used to select the optimal model.

Theorem 4.1 *Let* $P_D^0 = n\mathbf{tr}\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right]$, $P_D = P_D^0 + p$, $P_D^a = 3P_D^0 - n^2\mathbf{tr}\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right]$. *When the misspecification is ignored, we denote the first information criterion as*

$$IC = -2\log p(\mathbf{y}|\bar{\theta}) + P_D.$$

When the misspecification is taken into account and Müller's posterior distribution is used, we denote the second information criterion as

$$IC^a = -2\log p(\mathbf{y}|\bar{\theta}) + P_D^a.$$

Under Assumptions 1-7, we have,

$$-2r_{2n} = \int IC \times p(\mathbf{y}|D)d\mathbf{y} + o(1), \text{ i.e., } E_D(IC) = -2r_{2n} + o(1),$$
$$-2r^a_{2n} = \int IC^a \times p(\mathbf{y}|D)d\mathbf{y} + o(1), \text{ i.e., } E_D(IC^a) = -2r^a_{2n} + o(1).$$

Remark 4.1 *In the new criteria, $-2\log p(\mathbf{y}|\bar{\theta})$ can be understood as a Bayesian measure of fit, while P_D and P^a_D measure the model complexity. This feature of trade-off between the goodness of fit of the model and the complexity of the model is shared by other criteria.*

Remark 4.2 *In the iid case, note that $P^0_D = n\mathbf{tr}\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right] = \hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta}) + o_p(1)$. Hence, from Theorem 3.1, we have*

$$IC = -2\log p(\mathbf{y}|\bar{\theta}) + \hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta}) + p + o_p(1).$$

This compares to Akaike (1973)'s AIC, $-2\log p(\mathbf{y}|\hat{\theta}) + 2p$, and to Takeuchi (1976)'s TIC, $-2\log p(\mathbf{y}|\hat{\theta}) + 2\hat{\mathbf{J}}(\hat{\theta})\hat{\mathbf{I}}^{-1}(\hat{\theta})$.

Remark 4.3 *When the model is correctly specified, $P^0_D = \hat{\mathbf{J}}(\bar{\theta})\hat{\mathbf{I}}^{-1}(\bar{\theta}) + o_p(1) = p + o_p(1)$, $P_D = 2p + o_p(1)$, and $P^a_D = 2p + o_p(1)$. Since $AIC = -2\log p(\mathbf{y}|\hat{\theta}) + 2p$ and $DIC = AIC + o_p(1)$, we have,*

$$IC^a = IC + o_p(1) = TIC + o_p(1) = AIC + o_p(1) = DIC + o_p(1). \tag{10}$$

Equation (10) suggests that when a model is correctly specified, the two newly proposed information criteria, IC and IC^a, are asymptotically equivalent to each other. Furthermore, when a model is correctly specified, they are asymptotically equivalent to AIC, TIC and DIC.

Remark 4.4 *Similar to TIC, IC^a works for both correctly specified models and misspecified models. However, compared to TIC, IC^a is easier to compute as we do not need to invert $\hat{\mathbf{I}}(\bar{\theta})$. This advantage is especially important when the dimension of θ is high.*

Let the optimal decision under risk $r_n(d_j)$ and $r^a_n(d_j)$ be j^* and j^{a*}, respectively. By construction and by Theorem 4.1,

$$\lim_{n\to+\infty} r^a_n(d_{j^{a*}}) \leq \lim_{n\to+\infty} r^a_n(d_{j^*}) \leq \lim_{n\to+\infty} r_n(d_{j^*}).$$

Therefore, the risk associated with Müller's posterior cannot be lower (weakly) than that with the true posterior of the misspecified model.

Theorem 4.2 *Under Assumptions 1-7, if* $\mathbf{J}(\theta_0) \neq \mathbf{I}(\theta_0)$ *for model* j^* *and* j^{a*},

$$\lim_{n\to+\infty} r_n^a(d_{j^{a*}}) < \lim_{n\to+\infty} r_n(d_{j^*}).$$

When the models are misspecified and $\mathbf{J}(\theta_0) \neq \mathbf{I}(\theta_0)$, Theorem 4.2 suggests that the risk associated with Müller's posterior is strictly lower than that with the posterior implied by the misspecified model. This reduction in risk can be achieved in two ways. First, a different optimal model may be selected by the new risk function. Second, the same optimal model can be selected but the risk based on Müller's posterior is strictly smaller than that based on the original posterior. While we have the sufficient condition for the strict inequality for the risk functions, in general, unfortunately, it is difficult to give a sufficient condition under which the new risk selects a different optimal model (i.e., $j^{a*} \neq j^*$).

5 Conclusion

In this paper, we extend the idea and the results of Müller (2013) from Bayesian estimation to Bayesian model selection when candidate models are misspecified. It is shown that under model misspecification, on the basis of KL divergence between the correct predictive distribution and the model implied predictive distribution, Müller's posterior leads to a lower risk than that of the original posterior implied by the misspecified model. Two new model selection criteria are proposed. They are asymptotically equivalent to AIC, TIC and DIC when the model is correctly specified. When the candidate models are misspecified, the optimal model selected by Müller's posterior may be different from that based on the original posterior. Relative to TIC, our new information criteria are easier to compute.

6 Appendix

6.1 Appendix 1: Proof of Lemma 3.1

Using the first order Taylor expansion, for $t = 1, 2, \cdots, n$, we have

$$\begin{aligned}
& s_{n+t}(\tilde{\theta}_{n+t}) = s_{n+t}(\theta_0) + h_{n+t}(\theta_{n+t,0})(\tilde{\theta}_{n+t} - \theta_0) \\
& \sup_{t\leq n} ||s_{n+t}(\tilde{\theta}_{n+t})|| = \sup_{t\leq n} || \left[s_{n+t}(\theta_0) + h_{n+t}(\theta_{n+t,0})(\theta_{n+t} - \theta_0)\right] || \\
\leq\ & \sup_{t\leq n} ||s_{n+t}(\theta_0)|| + \sup_{t\leq n} ||h_{n+t}(\theta_{n+t,0})(\tilde{\theta}_{n+t} - \theta_0)|| \\
\leq\ & \sup_{t\leq n} ||s_{n+t}(\theta_0)|| + \left[\sup_{t\leq n} ||\frac{1}{\sqrt{n}} h_{n+t}(\theta_{n+t,0})||\right] \left[\sup_{t\leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)||\right],
\end{aligned}$$

where $\theta_{n+t,0}$ lies on the segment between $\tilde{\theta}_{n+t}$ and θ_0. Since θ_{n+t} is the consistent estimator of θ_0, there exists a real sequence $k_{n+t} \to 0$ such that $||\theta_{n+t} - \theta_0|| \leq k_{n+t}$ for enough large

n and $||\theta_{n+t,0} - \theta_0|| \leq k_{n+t}$. Using the regularity conditions and $\theta_{n+t,0}$ is dependent on $\mathbf{y}$, we have

$$\begin{aligned}
& \sqrt{n}(\tilde{\theta}_{n+t} - \theta_0) = O_p(1), \sup_{t\leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)|| = O_p(1), \\
& \sup_{t\leq n} \frac{1}{\sqrt{n+t}} ||s_{n+t}(\theta_0)|| \leq \sup_{t\leq n} \frac{1}{\sqrt{n+t}} \sup_{i\leq n+t} ||s_i(\theta_0)|| \\
\leq \quad & \frac{1}{\sqrt{n}} \sup_{t\leq n} \sup_{i\leq 2n} ||s_i(\theta_0)|| = \frac{1}{\sqrt{n}} o_p(\sqrt{2n}) = o_p(1), \\
& \sup_{t\leq n} ||\frac{1}{n+t} h_{n+t}(\theta_{n+t,0})|| \leq \frac{1}{n} \sup_{t\leq n} ||h_{n+t}(\theta_{n+t,0})|| \\
\leq \quad & \frac{1}{n} \sup_{||\theta-\theta_0||\leq k_{2n}} ||h_{2n}(\theta)|| = \frac{1}{n} o_p(2n) = o_p(1).
\end{aligned}$$

Hence, we get

$$\begin{aligned}
& \sup_{t\leq n} \frac{1}{\sqrt{n+t}} ||s_{n+t}(\theta_{n+t})|| \\
\leq \quad & \sup_{t\leq n} \frac{1}{\sqrt{n+t}} ||s_{n+t}(\theta_0)|| + \sup_{t\leq n} ||\frac{1}{\sqrt{n(n+t)}} h_{n+t}(\widetilde{\theta}_{n+t,0})|| \sup_{t\leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)|| \\
\leq \quad & \sup_{t\leq n} \frac{1}{\sqrt{n+t}} ||s_{n+t}(\theta_0)|| + \sup_{t\leq n} ||\frac{\sqrt{2}}{n+t} h_{n+t}(\widetilde{\theta}_{n+t,0})|| \sup_{t\leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)|| \\
= \quad & o_p(1) + \sqrt{2} o_p(1) O_p(1) = o_p(1), \\
& \sup_{t\leq n} \left[\frac{1}{n+t} ||s_{n+t}(\theta_{n+t}) s_{n+t}(\tilde{\theta}_{n+t})|| \right] \leq \left\{ \sup_{t\leq n} \left[\frac{1}{\sqrt{n+t}} ||s_{n+t}(\theta_{n+t})|| \right] \right\}^2 = o_p(1)^2 = o_p(1).
\end{aligned}$$

Further, we can show that

$$\begin{aligned}
& ||\frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t})|| - \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\theta_0)|| || = ||\frac{1}{n} \sum_{t=1}^{n} [||h_{n+t}(\theta_{n+t})|| - ||h_{n+t}(\theta_0)||] || \\
\leq \quad & \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t}) - h_{n+t}(\theta_0)|| \\
\leq \quad & \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t}) - h_{n+t}(\theta_0)|| + \frac{1}{n} \sum_{t=1}^{n} ||h_t(\tilde{\theta}_t) - h_t(\theta_0)|| \\
\leq \quad & 2 \sup_{||\theta_t-\theta_0||\leq k_{2n}} \left\{ \frac{1}{2n} \left[\sum_{t=1}^{2n} ||h_t(\theta_t) - h_t(\theta_0)|| \right] \right\} = o_p(1) \\
& \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t})|| = \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\theta_0)|| + o_p(1) = \frac{1}{n} \sum_{t=1}^{2n} ||h_t(\theta_0)|| - \frac{1}{n} \sum_{t=1}^{n} ||h_t(\theta_0)|| + o_p(1) \\
= \quad & 2 \frac{1}{2n} \sum_{t=1}^{2n} ||h_t(\theta_0)|| - \frac{1}{n} \sum_{t=1}^{n} ||h_t(\theta_0)|| + o_p(1) = 2O_p(1) - O_p(1) + o_p(1) = O_p(1).
\end{aligned}$$

Similarly, we can show that

$$
\begin{aligned}
& ||\frac{1}{n}\sum_{t=1}^{n} h_{n+t}(\tilde{\theta}_{n+t}) - \frac{1}{n}\sum_{t=1}^{n} h_{n+t}(\theta_0)|| = ||\frac{1}{n}\sum_{t=1}^{n}\left[h_{n+t}(\theta_{n+t}) - h_{n+t}(\theta_0)\right]|| \\
\leq \quad & \frac{1}{n}\sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t}) - h_{n+t}(\theta_0)|| \\
\leq \quad & 2 \sup_{||\theta-\theta||\leq k_{2n}} \left\{ \frac{1}{2n}\left[\sum_{t=1}^{2n} ||h_t(\theta_t) - h_t(\theta_0)||\right]\right\} = o_p(1) \\
& \frac{1}{n}\sum_{t=1}^{n} h_{n+t}(\tilde{\theta}_{n+t}) = \frac{1}{n}\sum_{t=1}^{n} h_t(\theta_0) + o_p(1) = \frac{1}{n}\sum_{t=1}^{2n} h_t(\theta_0) - \frac{1}{n}\sum_{t=1}^{n} h_t(\theta_0) + o_p(1) \\
= \quad & 2\frac{1}{2n}\sum_{t=1}^{2n} h_t(\theta_0) - \frac{1}{n}\sum_{t=1}^{n} h_t(\theta_0) + o_p(1) = -2\mathbf{I}(\theta_0) + \mathbf{I}(\theta_0) + o_p(1) = -\mathbf{I}(\theta_0) + o_p(1).
\end{aligned}
$$

Finally, we get

$$
\begin{aligned}
& \frac{1}{n}||\sum_{t=1}^{n} s_{n+t}(\tilde{\theta}_{n+t})s_{n+t}(\tilde{\theta}_{n+t})' - \sum_{t=1}^{n} s_{n+t}(\theta_0)s_{n+t}(\theta_0)'|| \\
= \quad & \frac{1}{n}||\sum_{t=1}^{n}\left[s_{n+t}(\theta_{n+t})s_{n+t}(\tilde{\theta}_{n+t})' - s_{n+t}(\theta_0)s_{n+t}(\theta_0)'\right]|| \\
\leq \quad & \frac{1}{n}\sum_{t=1}^{n} ||2h_{n+t}(\tilde{\theta}_{n+t,0})(\tilde{\theta}_{n+t} - \theta_0)s'_{n+t}(\theta_0)|| \\
& + \frac{1}{n}\sum_{t=1}^{n} ||h_{n+t}(\tilde{\theta}_{n+t,0})(\tilde{\theta}_{n+t} - \theta_0)(\tilde{\theta}_{n+t} - \theta_0)'h'_{n+t}(\theta_{n+t,0})|| \\
\leq \quad & \left[\frac{1}{n}\sum_{t=1}^{n} ||h_{n+t}(\theta_{n+t,0})||\right]\left[2||(\tilde{\theta}_{n+t} - \theta_0)|| \sup_{t\leq n} ||s'_{n+t}(\theta_0)||\right] \\
& + \left[\frac{1}{n}\sum_{t=1}^{n} ||h_{n+t}(\theta_{n+t,0})||\right]\left[||(\tilde{\theta}_{n+t} - \theta_0)||^2 \sup_{t\leq n} ||h_{n+t}(\theta_{n+t,0})||\right] \\
= \quad & 2O_p(1)O_p\left(n^{-1/2}\right)o_p\left(n^{1/2}\right) + O_p(1)O_p\left(n^{-1}\right)o_p\left(n\right) = o_p(1). \\
& \frac{1}{n}\sum_{t=1}^{n} s_{n+t}(\tilde{\theta}_{n+t})s_{n+t}(\tilde{\theta}_{n+t})' = \frac{1}{n}\sum_{t=1}^{n} s_{n+t}(\theta_0)s_{n+t}(\theta_0)' + o_p(1) \\
= \quad & 2\frac{1}{2n}\sum_{t=1}^{2n} s_t(\theta_0)s_t(\theta_0)' - \frac{1}{n}\sum_{t=1}^{n} s_t(\theta_0)s_t(\theta_0)' + o_p(1) \\
= \quad & 2\mathbf{J}(\theta_0) - \mathbf{J}(\theta_0) + o_p(1) = \mathbf{J}(\theta_0) + o_p(1). \\
& \frac{1}{n}\sum_{t=1}^{n} ||s_{n+t}(\tilde{\theta}_{n+t})s_{n+t}(\tilde{\theta}_{n+t})'|| = \mathbf{tr}\left[\frac{1}{n}\sum_{t=1}^{n} s_{n+t}(\tilde{\theta}_{n+t})s_{n+t}(\tilde{\theta}_{n+t})'\right] = O_p(1).
\end{aligned}
$$

6.2 Appendix 2: Proof of Lemma 3.2

Based on the standard ML theory and Lemma 3.1, we know that $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$, $\overline{\theta} - \widehat{\theta} = o_p(n^{-1/2})$. Hence, we have

$$\overline{\theta} - \theta_0 = \overline{\theta} - \hat{\theta} + \widehat{\theta} - \theta_0 = o_p(n^{-1/2}) + O_p(n^{-1/2}) = O_p(n^{-1/2}).$$

Theorem 3 of Müller (2013) implies that

$$\frac{1}{n} L_n^{(2)}(\hat{\theta}) = \frac{1}{n} \sum_{t=1}^{n} h_t(\hat{\theta}) = -\mathbf{I}^{-1}(\theta_0) + o_p(1).$$

Then, we have

$$\begin{aligned}
& E\left[(\theta - \theta_0)(\theta - \theta_0)'|\mathbf{y}\right] = \int \left[(\theta - \theta_0)(\theta - \theta_0)'\right] p(\theta|\mathbf{y}) d\theta \\
= & \int \left[(\theta - \hat{\theta})(\theta - \hat{\theta})' + 2(\theta - \hat{\theta})(\widehat{\theta} - \theta_0)' + (\hat{\theta} - \theta_0)(\widehat{\theta} - \theta_0)'\right] p(\theta|\mathbf{y}) d\theta \\
= & \int \left[(\theta - \hat{\theta})(\theta - \hat{\theta})'\right] p(\theta|\mathbf{y}) d\theta + 2(\overline{\theta} - \widehat{\theta})(\hat{\theta} - \theta_0)' + (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \\
= & -L_n^{-(2)}(\hat{\theta}) + 2(\overline{\theta} - \hat{\theta})(\hat{\theta} - \theta_0)' + (\widehat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \\
= & -L_n^{-(2)}(\hat{\theta}) + 2(\overline{\theta} - \hat{\theta})(\hat{\theta} - \theta_0)' + (\widehat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \\
= & -L_n^{-(2)}(\hat{\theta}) + 2o_p(n^{-1/2})O_p(n^{-1/2}) + (\hat{\theta} - \theta_0)(\hat{\theta} - \theta_0)' \\
= & \frac{1}{n}\left[\mathbf{I}^{-1}(\theta_0) + \mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(n^{-1}).
\end{aligned}$$

Furthermore, we can show that

$$\begin{aligned}
& V(\overline{\theta}) = E\left[(\theta - \overline{\theta})(\theta - \overline{\theta})'|\mathbf{y}\right] = \int \left[(\theta - \overline{\theta})(\theta - \overline{\theta})'\right] p(\theta|\mathbf{y}) d\theta \\
= & \int \left[(\theta - \hat{\theta})(\theta - \hat{\theta})' + 2(\theta - \hat{\theta})(\widehat{\theta} - \overline{\theta})' + (\hat{\theta} - \overline{\theta})(\widehat{\theta} - \overline{\theta})'\right] p(\theta|\mathbf{y}) d\theta \\
= & \int \left[(\theta - \hat{\theta})(\theta - \hat{\theta})'\right] p(\theta|\mathbf{y}) d\theta + 2(\overline{\theta} - \widehat{\theta})(\hat{\theta} - \overline{\theta})' + (\hat{\theta} - \overline{\theta})(\hat{\theta} - \overline{\theta})' \\
= & E\left[(\theta - \hat{\theta})(\theta - \hat{\theta})'|\mathbf{y}\right] - (\hat{\theta} - \overline{\theta})(\hat{\theta} - \overline{\theta})' \\
= & V(\hat{\theta}) + o_p(n^{-1/2})o_p(n^{-1/2}) = V(\widehat{\theta}) + o_p(n^{-1}).
\end{aligned}$$

Hence,

$$\begin{aligned}
& V(\overline{\theta}) = V(\hat{\theta}) + o_p(n^{-1}) = \frac{1}{n}[-\frac{1}{n} L_n^{(2)}(\widehat{\theta})]^{-1} + o_p(n^{-1}) \\
= & \frac{1}{n}\mathbf{I}^{-1}(\theta_0) + \frac{1}{n} o_p(1) + o_p(n^{-1}) = \frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(n^{-1}).
\end{aligned}$$

Similar to Theorem 3 of Müller (2013), we can show that

$$s_t(\overline{\theta}) = s_t(\widehat{\theta}) + h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \hat{\theta}),$$

where $\widetilde{\theta}$ lies on the segment between $\widehat{\theta}$ and $\overline{\theta}$, and that

$$\begin{aligned}
&\sum_{t=1}^{n} s_t(\overline{\theta}) s_t(\overline{\theta})' = \sum_{t=1}^{n} \left\{ [s_t(\widehat{\theta}) + h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \hat{\theta})][s_t(\hat{\theta}) + h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \widehat{\theta})]' \right\} \\
= \quad & \sum_{t=1}^{n} \left\{ s_t(\hat{\theta}) s_t(\widehat{\theta})' + 2h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \widehat{\theta}) s_t'(\hat{\theta}) + h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \widehat{\theta})(\overline{\theta} - \widehat{\theta})' h_t'(\widetilde{\theta}_{n+t}) \right\}.
\end{aligned}$$

Because $\hat{\theta}$ and $\overline{\theta}$ are both consistent estimators of θ_0, on the basis of Assumption 5, we can find some large n and null sequence k_n to make $\widetilde{\theta}_{n+t}$ and $\widehat{\theta}$ which both lie in $\{\theta : ||\theta - \theta_0|| \leq k_n\}$. Using Lemma 3.1, we can show that

$$\begin{aligned}
& \frac{1}{n} \sum_{t=1}^{n} ||h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \hat{\theta})' s_t'(\widetilde{\theta}_{n+t})|| \\
\leq \quad & \left[\frac{1}{n} \sum_{t=1}^{n} ||h_t(\widetilde{\theta}_{n+t})|| \right] \left[||(\overline{\theta} - \widehat{\theta})|| \right] \left[\sup_{t \leq n} ||s_t(\widehat{\theta})|| \right] \\
= \quad & O_p(1) O_p(n^{-1/2}) o_p(n^{1/2}) = o_p(1). \\
& \frac{1}{n} \sum_{t=1}^{n} ||h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \hat{\theta})(\overline{\theta} - \hat{\theta})' h_t'(\widetilde{\theta}_{n+t})|| \\
\leq \quad & \left[\frac{1}{n} \sum_{t=1}^{n} ||h_t(\widetilde{\theta}_{n+t})(\overline{\theta} - \widehat{\theta})(\overline{\theta} - \hat{\theta})'|| \right] \left[\sup_{t \leq n} ||h_t'(\widetilde{\theta}_{n+t})|| \right] \\
\leq \quad & \left[\frac{1}{n} \sum_{t=1}^{n} ||h_t(\widetilde{\theta}_{n+t})|| \right] \left[||(\overline{\theta} - \widehat{\theta})(\overline{\theta} - \widehat{\theta})'|| \right] \left[\sup_{t \leq n} ||h_t(\widetilde{\theta}_{n+t})|| \right] \\
= \quad & O_p(1) o_p(n^{-1}) o_p(n) = o_p(1).
\end{aligned}$$

and

$$\begin{aligned}
& \frac{1}{n} || \sum_{t=1}^{n} s_{n+t}(\overline{\theta}) s_{n+t}(\overline{\theta})' - \sum_{t=1}^{n} s_{n+t}(\widehat{\theta}) s_{n+t}(\hat{\theta})' || \\
= \quad & \frac{1}{n} || \sum_{t=1}^{n} \left[s_{n+t}(\overline{\theta}) s_{n+t}(\overline{\theta})' - s_{n+t}(\widehat{\theta}) s_{n+t}(\hat{\theta})' \right] || \\
\leq \quad & \frac{1}{n} \sum_{t=1}^{n} ||2h_{n+t}(\widetilde{\theta}_{n+t})(\overline{\theta} - \widehat{\theta}) s_{n+t}'(\hat{\theta})|| + \frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\widetilde{\theta}_{n+t})(\overline{\theta} - \hat{\theta})(\overline{\theta} - \hat{\theta})' h_{n+t}'(\widetilde{\theta}_{n+t})|| \\
\leq \quad & \left[\frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\widetilde{\theta}_{n+t})|| \right] \left[2||(\overline{\theta} - \widehat{\theta})|| \sup_{t \leq n} ||s_{n+t}'(\widehat{\theta})|| \right] \\
& + \left[\frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\widetilde{\theta}_{n+t})|| \right] \left[||(\overline{\theta} - \widehat{\theta})||^2 \sup_{t \leq n} ||h_{n+t}(\widetilde{\theta}_{n+t})|| \right] \\
= \quad & o_p(1).
\end{aligned}$$

Hence, we have

$$\frac{1}{n}\sum_{t=1}^{n} s_t(\overline{\theta})s_t(\overline{\theta})' = \frac{1}{n}\sum_{t=1}^{n} s_t(\widehat{\theta})s_t(\hat{\theta})' + o_p(1).$$

Based on Assumptions 1-6, according to Müller (2013), it can be shown that

$$\begin{aligned}
&\hat{\mathbf{I}}(\hat{\theta}) = -\frac{1}{n}L_n^{(2)}(\hat{\theta}) = -\frac{1}{2}\sum_{t=1}^{n} h_t(\widehat{\theta}) = \mathbf{I}(\theta_0) + o_p(1) \\
&\hat{\mathbf{J}}(\overline{\theta}) = \frac{1}{n}\sum_{t=1}^{n} s_t(\overline{\theta})s_t(\overline{\theta})' = \frac{1}{n}\sum_{t=1}^{n} s_t(\widehat{\theta})s_t(\hat{\theta})' + o_p(1) \\
= \quad & \frac{1}{n}\sum_{t=1}^{n} s_t(\theta_0)s_t(\theta_0)' + o_p(1) = \mathbf{J}(\theta_0) + o_p(1).
\end{aligned}$$

Furthermore, we can show that

$$\begin{aligned}
&V(\overline{\theta}) = V(\hat{\theta}) + o_p(n^{-1}) = [-L_n^{(2)}(\hat{\theta})]^{-1} + o_p(n^{-1}) \\
= \quad & \frac{1}{n}[-\frac{1}{n}L_n^{(2)}(\hat{\theta})]^{-1} + o_p(n^{-1}) = \frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(n^{-1}).
\end{aligned}$$

Hence, we have

$$\begin{aligned}
&\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) = n^2\left[\frac{1}{n}\mathbf{I}^{-1}(\theta_0)\right]\mathbf{J}(\theta_0)\left[\frac{1}{n}\mathbf{I}^{-1}(\theta_0)\right] \\
= \quad & n^2\left[V(\overline{\theta}) + o_p(n^{-1})\right]\left[\hat{\mathbf{J}}(\overline{\theta}) + o_p(1)\right]\left[V(\overline{\theta}) + o_p(n^{-1})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta}) + V(\overline{\theta})o_p(1) + o_p(n^{-1})\hat{\mathbf{J}}(\overline{\theta}) + o_p(n^{-1})\right]\left[V(\overline{\theta}) + o_p(n^{-1})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta}) + O_p(n^{-1})o_p(1) + o_p(n^{-1})O_p(1) + o_p(n^{-1})\right]\left[V(\overline{\theta}) + o_p(n^{-1})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta}) + o_p(n^{-1})\right]\left[V(\overline{\theta}) + o_p(n^{-1})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})o_p(n^{-1}) + V(\overline{\theta})o_p(n^{-1}) + o_p(n^{-2})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + O_p(n^{-1})O_p(1)o_p(n^{-1}) + O_p(n^{-1})o_p(n^{-1}) + o_p(n^{-2})\right] \\
= \quad & n^2\left[V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + o_p(n^{-2})\right] \\
= \quad & n^2V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + o_p(1),
\end{aligned}$$

and

$$\begin{aligned}
&E\left[(\theta-\theta_0)(\theta-\theta_0)'|\mathbf{y}\right] = \int\left[(\theta-\theta_0)(\theta-\theta_0)'\right]p(\theta|\mathbf{y})d\theta \\
= \quad & \frac{1}{n}\left[\mathbf{I}^{-1}(\theta_0) + \mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(n^{-1}) \\
= \quad & V(\overline{\theta}) + \frac{1}{n}\left[n^2V(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + o_p(1)\right] + o_p(n^{-1}) \\
= \quad & V(\overline{\theta}) + nV(\overline{\theta})\hat{\mathbf{J}}(\overline{\theta})V(\overline{\theta}) + o_p(n^{-1}).
\end{aligned}$$

6.3 Appendix 3: Proof of Theorem 3.1

Considering the loss function

$$\prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta)p(\theta|\mathbf{y})d\theta\right],$$

and applying the Taylor expansion to $p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta)$, we get

$$\begin{aligned}
&p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta) = p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0) + \frac{\partial p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}{\theta}(\theta-\theta_0)\\
&+\frac{1}{2}(\theta-\theta_0)'\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_{n+t})}{\theta\theta'}(\theta-\theta_0),
\end{aligned}$$

where $\widetilde{\theta}_{n+t}$ lies on the segment between θ and θ_0. Let $a_{n+t} = \frac{p(\mathbf{y}_{n+t}|\mathbf{y},\theta_{n+t})}{p(\mathbf{y}_{n+t}|\mathbf{y},\theta_0)}$ and we have

$$\begin{aligned}
&p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta) = p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0) + \frac{\partial p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}{\theta}(\theta-\theta_0)\\
&+\frac{1}{2}(\theta-\theta_0)'\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_{n+t})}{\theta\theta'}(\theta-\theta_0)\\
=\;& p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\left[1+\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}\frac{\partial p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}{\theta}(\theta-\theta_0)\right]\\
&+\frac{1}{2}p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)a_{n+t}\left[(\theta-\theta_0)'\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\tilde{\theta}_{n+t})}\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\tilde{\theta}_{n+t})}{\theta\theta'}(\theta-\theta_0)\right]\\
=\;& p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\left[1+\frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}{\theta}(\theta-\theta_0)\right]\\
&+\frac{1}{2}p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)a_{n+t}\left[(\theta-\theta_0)'\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\tilde{\theta}_{n+t})}\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\tilde{\theta}_{n+t})}{\theta\theta'}(\theta-\theta_0)\right]\\
=\;& p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)q_{n+t},
\end{aligned}$$

where

$$\begin{aligned}
q_{n+t} &= q_{1,n+t} + \frac{1}{2}q_{2,n+t},\\
q_{1,n+t} &= \frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)}{\theta}(\theta-\theta_0),\\
q_{2,n+t} &= a_{n+t}\left[(\theta-\theta_0)'\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_{n+t})}\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\tilde{\theta}_{n+t})}{\theta\theta'}(\theta-\theta_0)\right].
\end{aligned}$$

We can further show that

$$\begin{aligned}
&\prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\left(1+q_{n+t}\right)p(\theta|\mathbf{y})d\theta\right]\\
=\;&\prod_{t=1}^{n}\left[p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\prod_{t=1}^{n}\left[\int\left(1+q_{n+t}\right)p(\theta|\mathbf{y})d\theta\right]\\
=\;&\prod_{t=1}^{n}\left[p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\prod_{t=1}^{n}\left(1+\bar{q}_{n+t}\right),
\end{aligned}$$

where

$$\begin{aligned}
\bar{q}_{n+t} &= \bar{q}_{1,n+t} + \frac{1}{2}\bar{q}_{2,n+t}, \\
\bar{q}_{1,n+t} &= \frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta_0)}{\theta}(\overline{\theta} - \theta_0), \\
\bar{q}_{2,n+t} &= a_{n+t}\left[\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \tilde{\theta}_{n+t})}\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \tilde{\theta}_{n+t})}{\theta\theta'}E\left[(\theta-\theta_0)(\theta-\theta_0)'|\mathbf{y}\right]\right],
\end{aligned}$$

and that

$$\begin{aligned}
&\frac{1}{p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta)}\frac{\partial^2 p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta)}{\theta\theta'} = \frac{\partial^2 \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta)}{\partial\theta\partial\theta'} \\
&+\frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta)}{\partial\theta}\frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1}, \theta)}{\partial\theta'} \\
= \;& \frac{\partial^2 \log p(\mathbf{y}^{n+t}|\theta)}{\partial\theta\partial\theta'} - \frac{\partial^2 \log p(\mathbf{y}^{n+t-1}|\theta)}{\partial\theta\partial\theta'} \\
&+\left[\frac{\partial \log p(\mathbf{y}^{n+t}|\theta)}{\partial\theta} - \frac{\partial \log p(\mathbf{y}^{n+t-1}|\theta)}{\partial\theta}\right]\left[\frac{\partial \log p(\mathbf{y}^{n+t}|\theta)}{\partial\theta} - \frac{\partial \log p(\mathbf{y}^{n+t-1}|\theta)}{\partial\theta}\right]' \\
= \;& h_{n+t}(\theta) + s_{n+t}(\theta)s'_{n+t}(\theta).
\end{aligned}$$

Then, we have

$$\begin{aligned}
&\bar{q}_{1,n+t} = s'_{n+t}(\theta_0)(\overline{\theta} - \theta_0), \\
&\bar{q}_{2,n+t} = a_{n+t}\left\{\left[h_{n+t}(\theta_{n+t}) + s_{n+t}(\tilde{\theta}_{n+t})s'_{n+1}(\tilde{\theta}_{n+t})\right]E\left[(\theta-\theta_0)(\theta-\theta_0)'|\mathbf{y}\right]\right\}.
\end{aligned}$$

Note that $\widetilde{\theta}_{n+t}$ lies on the segment between θ and θ_0. So, $\widetilde{\theta}_{n+t} - \theta_0 = O_p\left(n^{1/2}\right)$. Since $\widetilde{\theta}_{n+t}$ is dependent on $\mathbf{y}$, $\sup_{t\leq n}||\widetilde{\theta}_{n+t} - \theta_0|| = O_p\left(n^{1/2}\right)$. Therefore, we can get

$$\begin{aligned}
&\log p(\mathbf{y}_{n+t}|\mathbf{y}, \tilde{\theta}_{n+t}) = \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0) + \frac{\partial \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0)}{\theta}(\theta_{n+t} - \theta_0) \\
&+\frac{1}{2}(\theta_{n+t} - \theta_0)'\frac{\partial^2 \log p(\mathbf{y}_{n+t}|\mathbf{y}, \tilde{\theta}_{n+t,0})}{\theta\theta'}(\theta_{n+t} - \theta_0) \\
= \;& \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0) + s_{n+t}(\theta_0)(\theta_{n+t} - \theta_0) + \frac{1}{2}(\theta_{n+t} - \theta_0)'h_{n+t}(\tilde{\theta}_{n+t,0})(\theta_{n+t} - \theta_0),
\end{aligned}$$

where $\widetilde{\theta}_{n+t,0}$ lies on the segment between $\tilde{\theta}_{n+t}$ and θ_0. Following the proof of Lemma 3.1 and Assumptions, it can be shown that $\sup_{i\leq n+t}||h_i(\theta_{n+t,0})|| = o_p(n+t)$ and $\sup_{i\leq n+t}||s_i(\theta_0)|| =$

$o_p\left((n+t)^{1/2}\right)$, $t = 1, 2 \cdots, n$. Furthermore, we can show that

$$
\begin{aligned}
& || \log p(\mathbf{y}_{n+t}|\mathbf{y}, \tilde{\theta}_{n+t}) - \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0)|| \\
= \; & ||s_{n+t}(\theta_0)(\theta_{n+t} - \theta_0) + \frac{1}{2}(\theta_{n+t} - \theta_0)' h_{n+t}(\tilde{\theta}_{n+t,0})(\theta_{n+t} - \theta_0)|| \\
\leq \; & ||s_{n+t}(\theta_0)||||(\theta_{n+t} - \theta_0)|| + \frac{1}{2}||(\theta_{n+t} - \theta_0)'||||h_{n+t}(\tilde{\theta}_{n+t,0})||||(\theta_{n+t} - \theta_0)|| \\
\leq \; & \left[\sup_{i \leq n+t} ||s_i(\theta_0)||\right] \left[\sup_{t \leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)||\right] \frac{1}{\sqrt{n}} \\
& + \frac{1}{2}\left[\sup_{i \leq n+t} ||h_i(\theta_{n+t,0})||\right] \left[\sup_{t \leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)||\right]^2 \frac{1}{n}. \\
& \sup_{t \leq n} a_{n+t} = \sup_{t \leq n} \left\{\exp\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}, \tilde{\theta}_{n+t}) - \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0)\right]\right\} \\
\leq \; & \sup_{t \leq n} \left\{\exp\left[|| \log p(\mathbf{y}_{n+t}|\mathbf{y}, \tilde{\theta}_{n+t}) - \log p(\mathbf{y}_{n+t}|\mathbf{y}, \theta_0)||\right]\right\} \\
\leq \; & \exp\left\{\left[\sup_{t \leq n} \sup_{i \leq n+t} ||s_i(\theta_0)||\right] \left[\sup_{t \leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)||\right]\right\} \times \\
& \exp\left\{\frac{1}{2}\left[\sup_{t \leq n} \sup_{i \leq n+t} ||h_i(\theta_{n+t,0})||\right] \left[\sup_{t \leq n} ||\sqrt{n}(\theta_{n+t} - \theta_0)||^2\right]\right\} \\
= \; & \exp\left[o_p(\sqrt{2n})O_p(1)\frac{1}{\sqrt{n}} + o_p(2n)O_p(1)O_p(1)\frac{1}{n}\right] \\
= \; & \exp(o_p(1)) = 1 + o_p(1).
\end{aligned}
$$

Under Assumptions 1-7, we have

$$
\begin{aligned}
& \sup_{t \leq n} ||q1_{n+t}|| = \sup_{t \leq n} ||s_{n+t}(\theta_0)(\overline{\theta} - \theta_0)|| \\
\leq \; & \sup_{t \leq n} \left[||s_{n+t}(\theta_0)||||(\overline{\theta} - \theta_0)||\right] \\
\leq \; & \left[\sup_{t \leq n} ||\frac{1}{\sqrt{n}} s_{n+t}(\theta_0)||\right] ||\sqrt{n}(\overline{\theta} - \theta_0)|| \\
= \; & \left[\sup_{t \leq n} \frac{\sqrt{n+t}}{\sqrt{n}}\right] o_p(1)O_p(1) = o_p(1).
\end{aligned}
$$

Further, strict stationarity implies that

$$
\begin{aligned}
& \int s_{n+t}(\theta_0) p(\mathbf{y}_{n+t-q}, \cdots, \mathbf{y}_{n+t}|D) \prod_{i=0}^{q} d\mathbf{y}_{n+t-i} \\
& \sum_{t=1}^{n} \left[\int q1_{n+t} p(\mathbf{y}_{n+t-q}, \cdots, \mathbf{y}_{n+t}|D) \prod_{i=0}^{q} d\mathbf{y}_{n+t-i}\right] = 0.
\end{aligned}
$$

Following Lemma 3.2 and Lemma 3.1, we have

$$
\begin{aligned}
& \sup_{t\leq n} || \left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\tilde{\theta}_{n+t}) s'_{n+t}(\theta_{n+t}) \right] V(\theta_0) || \\
\leq & \left[\sup_{t\leq n} || \frac{1}{n+t} h_{n+t}(\theta_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
& + \left[\sup_{t\leq n} || \frac{1}{n+t} s_{n+t}(\theta_{n+t}) s'_{n+1}(\tilde{\theta}_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
\leq & \left\{ \sup_{t\leq n} || \frac{1}{n+t} h_{n+t}(\theta_{n+t}) || \right\} \{ ||2nV(\theta_0)|| \} \\
& + \left\{ \sup_{t\leq n} || \frac{1}{n+t} s_{n+t}(\theta_{n+t}) s'_{n+t}(\tilde{\theta}_{n+t}) || \right\} \{ ||2nV(\theta_0)|| \} \\
= & \; o_p(1) 2n O_p(n^{-1}) + o_p(1) 2n O_p(n^{-1}) = o_p(1) \\
& \sup_{t\leq n} ||q2_{n+t}|| = \sup_{t\leq n} || \mathbf{tr} \left\{ a_{n+t} \left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t}) s'_{n+t}(\theta_{n+t}) \right] V(\theta_0) \right\} || \\
\leq & \left[\sup_{t\leq n} (a_{n+t}) \right] \sup_{t\leq n} || \mathbf{tr} \left\{ \left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t}) s'_{n+t}(\theta_{n+t}) \right] V(\theta_0) \right\} || \\
= & \left[1 + o_p(1) \right] o_p(1) = o_p(1).
\end{aligned}
$$

Further, we can show that

$$
\begin{aligned}
& \sum_{t=1}^{n} || \left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\tilde{\theta}_{n+t}) s'_{n+t}(\theta_{n+t}) \right] V(\theta_0) || \\
\leq & \left[\sum_{t=1}^{n} || \frac{1}{n+t} h_{n+t}(\theta_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
& + \left[\sum_{t=1}^{n} || \frac{1}{n+t} s_{n+t}(\theta_{n+t}) s'_{n+1}(\tilde{\theta}_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
\leq & \left[\sum_{t=1}^{n} || \frac{1}{n} h_{n+t}(\theta_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
& + \left[\sum_{t=1}^{n} || \frac{1}{n} s_{n+t}(\theta_{n+t}) s'_{n+1}(\tilde{\theta}_{n+t}) || \right] \left[\sup_{t\leq n} (n+t) \right] ||V(\theta_0)|| \\
\leq & \left[\frac{1}{n} \sum_{t=1}^{n} ||h_{n+t}(\theta_{n+t})|| \right] ||2nV(\theta_0)|| + \left[\frac{1}{n} \sum_{t=1}^{n} ||s_{n+t}(\theta_{n+t}) s'_{n+1}(\tilde{\theta}_{n+t})|| \right] [2nV(\theta_0)||] \\
= & \left[O_p(1) + o_p(1) \right] O_p(1) + \left[O_p(1) + o_p(1) \right] O_p(1) = O_p(1).
\end{aligned}
$$

For any matrix $\mathbf{A}$, using $|\mathbf{tr}(\mathbf{A})| \leq \sqrt{dim(\mathbf{A})}||\mathbf{A}||$, we have

$$
\begin{aligned}
& \sum_{t=1}^{n} ||q2_{n+t}|| = \sum_{t=1}^{n} ||\mathbf{tr}\left\{a_{n+t}\left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_0)s'_{n+t}(\theta_{n+t})\right]V(\theta_0)\right\}|| \\
\leq & \left[\sup_{t\leq n}|a_{n+t}|\right]\left\{\sum_{t=1}^{n} ||\mathbf{tr}\left\{\left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+1}(\theta_{n+t})\right]V(\theta_0)\right\}||\right\} \\
\leq & \left[\sup_{t\leq n}|a_{n+t}|\right]\left\{\sqrt{p}\sum_{t=1}^{n} ||\left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+1}(\theta_{n+t})\right]V(\theta_0)||\right\} \\
= & [1+o_p(1)]\,O_p(1) = O_p(1). \\
& ||\sum_{t=1}^{n}\left\{q2_{n+t} - \mathbf{tr}\left\{\left[h_{n+t}(\theta_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})\right]V(\theta_0)\right\}\right\}|| \\
= & ||\sum_{t=1}^{n}\mathbf{tr}\left\{[a_{n+t}-1]\left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})\right]V(\theta_0)\right\}|| \\
\leq & \sup_{t\leq n}|a_{n+t}-1|\left\{\sqrt{p}\sum_{t=1}^{n}|\left[h_{n+t}(\tilde{\theta}_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+t}(\theta_{n+t})\right]V(\theta_0)||\right\} \\
= & o_p(1)O_p(1) = o_p(1). \\
& \sum_{t=1}^{n} q2_{n+t} = \sum_{t=1}^{n}\mathbf{tr}\left\{\left[h_{n+t}(\theta_{n+t}) + s_{n+t}(\theta_{n+t})s'_{n+t}(\tilde{\theta}_{n+t})\right]V(\theta_0)\right\} + o_p(1) \\
= & [-\mathbf{I}(\theta_0) + \mathbf{J}(\theta_0) + o_p(1)]\,[\mathbf{F}(\theta_0) + o_p(1)] \\
= & [-\mathbf{I}(\theta_0) + \mathbf{J}(\theta_0)]\,\mathbf{F}(\theta_0) + o_p(1).
\end{aligned}
$$

Using Lemma 3.2, we can further show that

$$
\begin{aligned}
& \sum_{t=1}^{n} ||q1_{n+t}^2|| = \sum_{t=1}^{n} q1_{n+t}^2 = \sum_{t=1}^{n}(\overline{\theta}-\theta_0)'s_{n+t}(\theta_0)s'_{n+t}(\theta_0)(\overline{\theta}-\theta_0) \\
= & \mathbf{J}(\theta_0)\mathbf{F}(\theta_0) + o_p(1) \\
& ||\sum_{t=1}^{n}\left[q_{n+t}^2 - q1_{n+t}^2\right]|| = ||\sum_{t=1}^{n}\left\{\left[q1_{n+t} + \frac{1}{2}q2_{n+t}\right]^2 - q1_{n+t}^2\right\}|| \\
= & ||\sum_{t=1}^{n}\left[q1_{n+t}q2_{n+t} + \frac{1}{4}q2_{n+t}^2\right]|| \leq ||\sum_{t=1}^{n} q1_{n+t}q2_{n+t}|| + \frac{1}{4}||\sum_{t=1}^{n} q2_{n+t}^2|| \\
\leq & \sum_{t=1}^{n}||q2_{n+t}||\left[\sup_{t\leq n}||q1_{n+t}||\right] + \frac{1}{4}\sum_{t=1}^{n}||q2_{n+t}||\left[\sup_{t\leq n}||q2_{n+t}||\right] \\
= & O_p(1)o_p(1) + \frac{1}{4}O_p(1)o_p(1) = o_p(1). \\
& \sum_{t=1}^{n} q_{n+t}^2 = \sum_{t=1}^{n} q1_{n+t}^2 + o_p(1) = \sum_{t=1}^{n}\left[(\overline{\theta}-\theta_0)s_{n+t}(\theta_0)s'_{n+t}(\theta_0)(\overline{\theta}-\theta_0)\right] + o_p(1) \\
= & \mathbf{J}(\theta_0)\mathbf{F}(\theta_0) + o_p(1).
\end{aligned}
$$

Furthermore, we have

$$
\begin{aligned}
& ||\sum_{t=1}^{n} q_{n+t}^3|| = ||\sum_{t=1}^{n}\left[q1_{n+t} + \frac{1}{2}q2_{n+t}\right]^3|| \\
= \;& ||\sum_{t=1}^{n}\left[q1_{n+t}^3 + \frac{3}{2}q1_{n+t}^2 q2_{n+t} + \frac{3}{4}q1_{n+t}q2_{n+t}^2 + \frac{1}{8}q2_{n+t}^3\right]|| \\
\leq \;& ||\sum_{t=1}^{n} q1_{n+t}^3|| + \frac{3}{2}||\sum_{t=1}^{n} q1_{n+t}^2 q2_{n+t}|| + \frac{3}{4}\sum_{t=1}^{n}||q1_{n+t}q2_{n+t}^2|| + \frac{1}{8}||\sum_{t=1}^{n} q2_{n+t}^3|| \\
\leq \;& \left[\sum_{t=1}^{n}||q1_{n+t}^2||\right]\left[\sup_{t\leq n}||q1_{n+t}||\right] + \frac{3}{2}\left[||\sum_{t=1}^{n} q2_{n+t}||\right]\left[\sup_{t\leq n}||q1_{n+t}||\right]^2 \\
& + \frac{3}{4}\sum_{t=1}^{n}||q2_{n+t}||\left[\sup_{t\leq n}||q1_{n+t}||\right]\left[\sup_{t\leq n}||q2_{n+t}||\right] + \frac{1}{8}\left[\sum_{t=1}^{n}||q2_{n+t}||\right]||\left[\sup_{t\leq n}||q2_{n+t}||\right]^2 \\
= \;& O_p(1)o_p(1) + \frac{3}{2}O_p(1)o_p(1)^2 + \frac{3}{4}O_p(1)o_p(1)o_p(1) + \frac{1}{8}O_p(1)o_p(1) = o_p(1), \\
& \sum_{t=1}^{n} q_{n+t}^3 = o_p(1).
\end{aligned}
$$

Similarly, we can get that $||\sum_{t=1}^{n} q_{n+t}^j|| \leq ||\sum_{t=1}^{n} q_{n+t}^{j-1}||\left[\sup_{t\leq n}||q_{n+t}||\right] = o_p(1), j \geq 4$. Hence, we can obtain $\sum_{t=1}^{n} q_{n+t}^j = o_p\left(\sum_{t=1}^{n} q_{n+t}^3\right), j \geq 4$. For $x^3 = o_p(1)$, using $\log(1+x) = x - \frac{1}{2}x^2 + o_p(1)$, we get

$$
\begin{aligned}
& \int \log\left\{\prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\left(1+q_{n+t}\right)p(\theta|\mathbf{y})d\theta\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} \\
= \;& \int \log\left\{\prod_{t=1}^{n}\left[p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\prod_{t=1}^{n}\left(1+\bar{q}_{n+t}\right)\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} \\
= \;& \int \left\{\sum_{t=1}^{n}\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right] + \sum_{t=1}^{n}\left[\log\left(1+\bar{q}_{n+t}\right)\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} \\
= \;& \int \left\{\sum_{t=1}^{n}\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right] + \sum_{t=1}^{n}\bar{q}_{n+t} - \frac{1}{2}\sum_{t=1}^{n}\bar{q}_{n+t}^2 + o_p(1)\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} \\
= \;& \int \left\{\sum_{t=1}^{n}\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} + \frac{1}{2}\mathbf{tr}\left\{\left[-\mathbf{I}(\theta_0) + \mathbf{J}(\theta_0)\right]\mathbf{F}(\theta_0)\right\} \\
& -\frac{1}{2}\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1) \\
= \;& \int \left\{\sum_{t=1}^{n}\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} - \frac{1}{2}p + o(1) \\
= \;& \int \log p(\mathbf{y}_{n+1:2n}|\mathbf{y},\theta_0)p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} - \frac{1}{2}p + o(1) \\
= \;& \int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)d\mathbf{y} - \frac{1}{2}p + o(1).
\end{aligned}
$$

Then, when the original posterior distribution $p(\theta|\mathbf{y})$ is replaced with Müller posterior distribution $p^a(\theta|\mathbf{y})$, we have

$$\begin{aligned} & V^a(\theta_0) = E^a\left[(\theta-\theta_0)(\theta-\theta_0)'|\mathbf{y}\right] = 2\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + o_p(1) \\ = \quad & \mathbf{F}^a(\theta_0) + o_p(1). \end{aligned}$$

Similar to the proof above, if $V(\theta_0)$ is replaced by $V^a(\theta_0)$, we have

$$\begin{aligned} & \int \log\left\{\prod_{t=1}^{n}\left[\int p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\,(1+q_{n+t})\,p(\theta|\mathbf{y})d\theta\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} \\ = \quad & \int\left\{\sum_{t=1}^{n}\left[\log p(\mathbf{y}_{n+t}|\mathbf{y}^{n+t-1},\theta_0)\right]\right\} p(\mathbf{y}_{1:2n}|D)d\mathbf{y}_{1:2n} + \frac{1}{2}\mathbf{tr}\left\{\left[-\mathbf{I}(\theta_0)+\mathbf{J}(\theta_0)\right]\mathbf{F}^a(\theta_0)\right\} \\ & -\frac{1}{2}\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1) \\ = \quad & \int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)d\mathbf{y} + \frac{1}{2}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) - 2\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1). \end{aligned}$$

Hence, we can get that

$$\begin{aligned} -2r_n \quad &= \quad -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)d\mathbf{y} + p + o(1), \\ -2r_n^a \quad &= \quad -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)d\mathbf{y} - \left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) - 2\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1). \end{aligned}$$

We can further show that

$$\begin{aligned} & -2r_{2n} - (-2r_{2n}^a) \\ = \quad & \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) - 2\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + p + o(1). \end{aligned}$$

Let $\mathbf{C}(\theta_0) = \mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)$ and we get

$$\begin{aligned} & \mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0))] - 2\mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)] + p \\ = \quad & \mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0))] - 2\mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)] + p \\ = \quad & \mathbf{tr}[\mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)] \\ & -2\mathbf{tr}[\mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)] + p \\ = \quad & \mathbf{tr}\left[\mathbf{C}(\theta_0)\mathbf{C}(\theta_0)\right] - 2\mathbf{tr}\left[\mathbf{C}(\theta_0)\right] + p \\ = \quad & (\delta_1^2+\delta_2^2+\cdots+\delta_p^2) - (\delta_1+\delta_2+\cdots+\delta_p) + p \\ = \quad & \sum_{i=1}^{p}(\delta_i^2 - 2\delta_i + 1) = \sum_{i=1}^{p}(\delta_i-1)^2 \geq 0, \end{aligned}$$

where $\delta_i, i = 1, 2, \cdots, p$, are the eigenvalue of $\mathbf{C}(\theta_0)$. Hence, we have

$$\begin{aligned} & \lim_{n\to+\infty}\left[-r_{2n}\right] \geq \lim_{n\to+\infty}\left[-r_{2n}^a\right], \\ & \lim_{n\to+\infty}\left[r_{1n}-r_{2n}\right] \geq \lim_{n\to+\infty}\left[r_{1n}-r_{2n}^a\right], \\ & \lim_{n\to+\infty} r_n \geq \lim_{n\to+\infty} r_n^a. \end{aligned}$$

6.4 Appendix 4: Proof of Theorem 4.1

Using the Taylor expansion, we can show that

$$\begin{aligned}
& \log p(\mathbf{y}|\theta_0) = \log p(\mathbf{y}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}|\hat{\theta})}{\partial \theta}(\theta_0 - \hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 \log p(\mathbf{y}|\hat{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 \log p(\mathbf{y}|\tilde{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) + \frac{1}{2}(\theta_0 - \hat{\theta})' \frac{\partial^2 \log p(\mathbf{y}|\hat{\theta})}{\partial \theta \partial \theta}(\theta_0 - \hat{\theta}) + o_p(1) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) + \frac{1}{2}\left[\sqrt{n}(\theta_0 - \hat{\theta})'\right]\left[\frac{1}{2}\sum_{t=1}^{n} h_t(\hat{\theta})\right]\left[\sqrt{n}(\theta_0 - \hat{\theta})\right] + o_p(1) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) + \frac{1}{2}\left[\sqrt{n}(\hat{\theta} - \theta_0)'\right]\left[\frac{1}{2}\sum_{t=1}^{n} h_t(\hat{\theta})\right]\left[\sqrt{n}(\hat{\theta} - \theta_0)\right] + o_p(1) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) - \frac{1}{2}\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(1).
\end{aligned}$$

$$\begin{aligned}
& \log p(\mathbf{y}|\bar{\theta}) = \log p(\mathbf{y}|\hat{\theta}) + \frac{\partial \log p(\mathbf{y}|\hat{\theta})}{\partial \theta}(\bar{\theta} - \hat{\theta}) + \frac{1}{2}(\bar{\theta} - \hat{\theta})' \frac{\partial^2 \log p(\mathbf{y}|\hat{\theta})}{\partial \theta \partial \theta}(\bar{\theta} - \hat{\theta}) + o_p(1) \\
= \;& \log p(\mathbf{y}|\hat{\theta}) + 0 + \frac{1}{2} o_p(n^{-1/2}) O_p(n) o_p(n^{-1/2}) + o_p(1) = \log p(\mathbf{y}|\hat{\theta}) + o_p(1) \\
= \;& \log p(\mathbf{y}|\theta_0) + \frac{1}{2}\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(1).
\end{aligned}$$

When the likelihood information dominates the prior information, we get

$$L_n^{(2)}(\theta) = -\mathbf{I}(\theta).$$

Using Lemma 3.1, we can show that

$$\begin{aligned}
& n\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right] = n\left[\mathbf{J}(\theta_0) + o_p(1)\right]\left[\frac{1}{n}\mathbf{I}^{-1}(\theta_0) + o_p(n^{-1})\right] \\
= \;& \left[\mathbf{J}(\theta_0) + o_p(1)\right]\left[\mathbf{I}^{-1}(\theta_0) + o_p(1)\right] \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + \mathbf{J}(\theta_0)o_p(1) + o_p(1)\mathbf{I}^{-1}(\theta_0) + o_p(1) \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + O_p(1)o_p(1) + O_p(1)o_p(1) + o_p(1) \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + o_p(1) \\
& n^2\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right] \\
= \;& \left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + o_p(1)\right]\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + o_p(1)\right] \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + 2\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)o_p(1)\right] + o_p(1) \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + 2\left[O_p(1)O_p(1)o_p(1)\right] + o_p(1) \\
= \;& \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + o_p(1).
\end{aligned}$$

Then, we can show that

$$\begin{aligned}
P_D &= P_D^0 + p = n\mathbf{tr}\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right] + p = \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + p + o_p(1) \\
P_D^a &= 3P_D^0 - n^2\mathbf{tr}\left[\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\hat{\mathbf{J}}(\bar{\theta})\mathbf{V}(\bar{\theta})\right] \\
&= 3\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] - \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(1).
\end{aligned}$$

Hence, according to Theorem **??**, we can get that

$$\begin{aligned}
&\int IC \times p(\mathbf{y}|D)\mathrm{d}\mathbf{y} = \int \left[-2\log p(\mathbf{y}|\bar{\theta}) + P_D\right] p(\mathbf{y}|D)\mathrm{d}\mathbf{y} \\
= \; & -2\int \log p(\mathbf{y}|\bar{\theta})p(\mathbf{y}|D)\mathrm{d}\mathbf{y} + \int P_D p(\mathbf{y}|D)\mathrm{d}\mathbf{y} \\
= \; & -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)\mathrm{d}\mathbf{y} - \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1) + \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + p + o(1) \\
= \; & -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)\mathrm{d}\mathbf{y} + p + o(1) \\
= \; & -2r_{2n} + o(1). \\
&\int IC^a \times p(\mathbf{y}|D)\mathrm{d}\mathbf{y} = \int \left[-2\log p(\mathbf{y}|\bar{\theta}) + P_D^a\right] p(\mathbf{y}|D)\mathrm{d}\mathbf{y} \\
= \; & -2\int \log p(\mathbf{y}|\bar{\theta})p(\mathbf{y}|D)\mathrm{d}\mathbf{y} + \int P_D^a p(\mathbf{y}|D)\mathrm{d}\mathbf{y} \\
= \; & -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)\mathrm{d}\mathbf{y} - \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1) \\
& +3\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] - \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o_p(1) \\
= \; & -2\int \log p(\mathbf{y}|\theta_0)p(\mathbf{y}|D)\mathrm{d}\mathbf{y} + 2\mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] - \mathbf{tr}\left[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] + o(1) \\
= \; & -2r_{2n}^a + o(1).
\end{aligned}$$

6.5 Appendix 5: Proof of Theorem 4.2

As shown in Theorem 3.1, for any statistical decision d, we can show that

$$\begin{aligned}
& -2r_{2n}(d) - (-2r_{2n}^a(d)) \\
= \; & -2\int \log p(\mathbf{y}|\bar{\theta})p(\mathbf{y}|D)d\mathbf{y} + \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + p \\
& +2\int \log p(\mathbf{y}|\bar{\theta})p(\mathbf{y}|D)d\mathbf{y} - \left[3\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) - \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\right] \\
= \; & \mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) - 2\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0) + p + o(1).
\end{aligned}$$

Hence, for the statistical decision d_j to choose j-th model, according to Theorem 3.1, we have

$$\begin{aligned}
&\lim_{n\to+\infty}\left[-r_{2n}(d_j)\right] \geq \lim_{n\to+\infty}\left[-r_{2n}^a(d_j)\right], \\
&\lim_{n\to+\infty}\left[r_{1n}(d_j) - r_{2n}(d_j)\right] \geq \lim_{n\to+\infty}\left[r_{1n}(d_j) - r_{2n}^a(d_j)\right], \\
&\lim_{n\to+\infty}\left[r_n(d_j)\right] \geq \lim_{n\to+\infty}\left[r_n^a(d_j)\right].
\end{aligned}$$

Let $\mathbf{C}(\theta_0) = \mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)$, similarly shown in Theorem 3.1, we know that

$$\begin{aligned} & \mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0))] - 2\mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)] + p \\ = & (\delta_1^2 + \delta_2^2 + \cdots + \delta_p^2) - (\delta_1 + \delta_2 + \cdots + \delta_p) + p \\ = & \sum_{i=1}^{p}(\delta_i^2 - 2\delta_i + 1) = \sum_{i=1}^{p}(\delta_i - 1)^2 \geq 0, \end{aligned}$$

where $\delta_i, i = 1, 2, \cdots, p$, are the eigenvalue of $\mathbf{C}(\theta_0)$. Hence, only and only if all the eigenvalues of $\mathbf{C}(\theta_0)$ are equal 1, the equality are true.

Since j^* and j^{a*} are the optimal decision under risk $r_n(d_j)$ and $r_n^a(d_j)$, we have

$$\lim_{n\to+\infty}[r^a(d_{j^{a*}})] \leq \lim_{n\to+\infty}[r^a(d_{j^*})] \leq \lim_{n\to+\infty}[r(d_{j^*})]\,.$$

If $\mathbf{J}(\theta_0) \neq \mathbf{I}(\theta_0)$, $\mathbf{C}(\theta_0) = \mathbf{I}^{-1/2}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1/2}(\theta_0)$ is not an identity matrix so that not all the eigenvalue of $\mathbf{C}(\theta_0)$ are equal one. Hence, we have

$$\mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0))] - 2\mathbf{tr}[\mathbf{J}(\theta_0)\mathbf{I}^{-1}(\theta_0)] + p > 0.$$

For model j^*, if $\mathbf{J}(\theta_0)\mathbf{I}(\theta_0)$ is not an identity matrix, we have

$$\lim_{n\to+\infty}[r_n^a(d_{j^{a*}})] \leq \lim_{n\to+\infty}[r_n^a(d_{j^*})] < \lim_{n\to+\infty}[r_n(d_{j^*})]\,.$$

Theorem 4.2 is proven.

References

Akaike, H. (1973) Information theory and the Maximum Likelihood Principle, in: B. N. Petrov & F. Csaki (Eds) *Second International Symposium on Information Theory* (Budapest: Akademiai Kiado)

Ferguson,T.S. (1973). A Bayesian Analysis of Some Nonparametric Problems. *Annals of Statistics*, **1(2)**, 209-230.

Gelfand, A. and Ghosh, S. (1998). Model Choice: A Minimum Posterior Predictive Loss Approach. *Biometrika*, **85**, 1-11

Ghosh, J.K. and Ramamoorthi, R.V. (2003). *Bayesian Nonparametrics.* Springer Publishing House.

Geweke, J.F. (2005). *Contemporary Bayesian Econometrics and Statistics*. Wiley.

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

Gourieroux, C., A. Monfort, A. Trognon, (1984b), Pseudo maximum likelihood methods: Applications to Poisson models. *Econometrica*, **52(3)**, 701–720.

Hansen, L. P. (1982). Large Sample Properties of Generalized Method of Moments Estimators. *Econometrica*, **50(4)**, 1029–1054.

Huber, P. J. (1967). The Behavior of Maximum Likelihood Estimates under Nonstandard Conditions. *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*,**vol. I**, 221-233.

Jensen,M.J. and Maheu, J.M. (2010). Bayesian Semiparametric Stochastic Volatility Modeling. *Journal of Econometrics*, **157(2)**, 306-316.

Jensen,M.J. and Maheu, J.M. (2013). Estimating a Semiparametric Asymmetric Stochastic Volatility Model with a Dirichlet Process Mixture. *Journal of Econometrics*, forthcoming.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factor. *Journal of the Americana Statistical Association*, **90**, 773-795.

Laud, P. W. and Ibrahim, J. G. (1995). Predictive Model Selection. *Journal of the Royal Statistical Society B*, **57**, 247-262.

Li, Y. and Yu, J. (2012). Bayesian Hypothesis Testing in Latent Variable Models. *Journal of Econometrics*, **166(2)**, 237-246.

Li, Y., Zeng, T. and Yu, J. (2012). A Robust Deviation Information Criterion for Latent Variable Models. *Working paper*, Singapore Management University.

Müller, U.K. (2013). Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix. *Econometrica*,**81(5)**,1805-1849.

Pericchi L.R. and Pérez M.E. (1994). Bayesian Robustness with more than one Sampling Model (with discussion). *Journal of Statistical Planning and Inference*, **40(2-3)**, 279-294.

Phillips, P.C.B. (1996) Econometric Model Determination, *Econometrica*, **64**, 763-812.

Schwarz, G. (1978) Estimating the Dimension of a Model, *Annals of Statistics*, **6**, 461–464.

Spiegelhalter, D., Best, N.G. and Carlin, B. and van der Linde, A. (2002). Bayesian Measures of Model Complexity and Fit. *Journal of the Royal Statistical Society B*, **64**, 583-639.

Takeuchi, K. (1976). Distribution of informational statistics and a criterion for model fitting. (In Japanse). Suri-Kagaku, **153**, 12-18.

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, **50(1)**, 1-25.