Submitted to *Econometrica*

RISK OF PREDICTIVE DISTRIBUTIONS AND BAYESIAN MODEL COMPARISON OF MISSPECIFIED MODELS

YONG LI
School of Economics, Renmin University of China

ZHOU WU
School of Economics, Zhejiang University

JUN YU
Faculty of Business Administration, University of Macau

TAO ZENG
School of Economics, Zhejiang University

Müller (2013, Econometrica, 81(5), 1805-1949) shows that Bayesian inference of parameters of interest in a misspecified model can reduce the asymptotic frequentist risk when the standard posterior is replaced with the sandwich posterior. In this paper, we extend the results in Müller (2013) to Bayesian model comparison. Bayesian model comparison of potentially misspecified models can be conducted in a predictive framework with three alternative predictive distributions, namely, the plug-in predictive distribution, the standard posterior predictive distribution, and the sandwich posterior predictive distribution of Müller (2013). Via the Kullback-Leibler (KL) loss function, it is shown that the sandwich posterior predictive distribution yields a lower asymptotic risk than the standard posterior predictive distribution. Moreover, we provide sufficient conditions under which the sandwich posterior predictive distribution yields a lower asymptotic risk than the plug-in predictive distribution. We then propose two new Bayesian penalized

Yong Li: gibbsli@163.com
Zhou Wu: wuzhou@zju.edu.cn
Jun Yu: junyu@um.edu.mo
Tao Zeng: ztzt6512@gmail.com

Li gratefully acknowledges the financial support of the National Natural Science Foundation of China (Grant Nos. 72273142, and 72394392). Zeng gratefully acknowledges the financial support from the National Natural Science Foundation of China (No. 72073121).

information criteria based on the last two predictive distributions to compare misspecified models and establish their relationship with some existing information criteria. The proposed new information criteria are illustrated in several empirical studies.

KEYWORDS: AIC, DIC, Information criterion, Model misspecification, Sandwich posterior.

1. INTRODUCTION

In many empirical studies, researchers frequently utilize simple parametric models. However, these models often lead to model misspecification. As George Box famously stated, "all models are wrong, but some are useful." When a model is misspecified, it can result in inefficient and even inconsistent estimation of parameters of interest. Moreover, likelihood-based statistical inferences – such as hypothesis testing and goodness-of-fit tests – are significantly impacted. Therefore, developing effective methods to address model misspecification is crucial.

White (1982) explored the consequences and detection of model misspecification in the context of maximum likelihood (ML) estimation and inference. He found that, within linear regression models, if the error distribution is misspecified and the normal distribution is incorrectly assumed for the likelihood function, the ML estimator (MLE) remains consistent and has an asymptotically normal distribution characterized by the so-called sandwich covariance matrix. Conversely, in the Bayesian framework, the standard posterior distribution is centered around the MLE and asymptotically follows a normal distribution, with its posterior variance converging to the Hessian information matrix. This indicates that standard posterior analysis does not provide adequate protection against model misspecification. In a significant contribution, Müller (2013) proposed conducting Bayesian analysis based on a sandwich posterior – an artificial Gaussian posterior centered at the MLE, with the sandwich covariance matrix as the posterior variance. He demonstrated that this approach yields Bayesian inference with lower asymptotic frequentist risk for parameters of interest.

Empirical researchers frequently face another critical statistical decision: model comparison. Notable studies on this topic include those by Granger et al. (1995), Phillips (1995), Phillips (1996), Hansen (2005), and Kadane and Lazar (2004). From a predictive perspective, several penalty-based information criteria have been developed for model comparison.

In the frequentist approach, two well-known criteria were proposed by Akaike (1974) and Takeuchi (1976). The former generally assumes that all candidate models encompass the true model or are good approximations of the data generating process (DGP), while the latter accommodates misspecified candidate models. Within the Bayesian framework, the Deviance Information Criterion (DIC) introduced by Spiegelhalter et al. (2002) is a commonly used penalty-based criterion. Li et al. (2025) provided a decision-theoretic explanation for DIC, asserting that it serves as the Bayesian counterpart to AIC. Furthermore, Li et al. (2020) developed a variant of DIC for comparing misspecified models. All these criteria are grounded in the Kullback-Leibler (KL) loss function and the plug-in predictive distribution.

In this paper, we focus on comparing alternative models that may be misspecified. Specifically, we aim to conduct a Bayesian comparison of these models based on their predictive performance. Similar to the existing criteria reviewed above, we utilize the KL function to define the risk. However, we consider three predictive distributions for the KL function calculation: the plug-in predictive distribution, the Bayesian predictive distribution derived from the standard posterior distribution, and the Bayesian predictive distribution based on the sandwich posterior distribution (hereafter referred to as the sandwich Bayesian predictive distribution). One of our main objectives is to evaluate the performance of these alternative predictive distributions and establish conditions under which their performance in terms of KL loss can be compared.

We investigate the theoretical properties of these three predictive distributions through asymptotic frequent risk analysis. Our findings indicate that the sandwich Bayesian predictive distribution exhibits lower frequentist risk than the standard posterior distribution. This result extends Müller (2013) findings to the context of model comparison. However, we generally cannot directly compare the frequentist risks associated with the plug-in predictive distribution and the Bayesian predictive distribution; this comparison depends on the degree of misspecification. We provide conditions under which such a comparison is possible.

Based on the frequentist risk analysis of the KL loss function derived from these predictive distributions, we propose two new penalty-based information criteria for model comparison. We demonstrate that these new criteria are asymptotically unbiased estimators of the risk between the corresponding predictive distributions and the DGP. Furthermore,

they are applicable for comparing misspecified models. We also establish the relationship between our proposed criteria and some existing ones, such as AIC and BIC. A significant advantage of our proposed information criteria is that they not only identify the optimal model but also the optimal predictive distribution, thereby offering more comprehensive insights from a predictive perspective compared to existing penalized information criteria.

The paper is organized as follows. Section 2 provides a brief review of the literature on statistical inferences regarding misspecified models. Section 3 examines the risk properties of the alternative predictive distributions. Section 4 introduces the new penalty-based information criteria for comparing misspecified models. Section 5 studies the performance of new information criteria by simulation experiments. Section 6 illustrates the application of the new methods. Finally, Section 7 concludes the paper. The Appendix contains the proofs of the two theorems presented, and the Online Appendix includes proofs of additional theoretical results.

2. STATISTICAL INFERENCE FOR MISSPECIFIED MODELS: A REVIEW

2.1. *MLE-based inference under model misspecification*

Let the observed data be $\mathbf{y} = (y_1, \cdots, y_n)$ with the DGP being $g(\mathbf{y})$. Consider a parametric model, denoted by $p(\mathbf{y}|\boldsymbol{\theta})$, which is used to fit the data, where $\boldsymbol{\theta}$ is the vector of parameters with P dimensions and $\boldsymbol{\theta} \in \boldsymbol{\Theta} \subseteq R^P$. In the literature, the KL divergence is used to measure the 'distance' between two distributions, say $g(\mathbf{y})$ and $p(\mathbf{y}|\boldsymbol{\theta})$, that is,

$$KL[g(\mathbf{y}), p(\mathbf{y}|\boldsymbol{\theta})] = \int g(\mathbf{y}) \ln \frac{g(\mathbf{y})}{p(\mathbf{y}|\boldsymbol{\theta})} d\mathbf{y} = E_{g(\mathbf{y})} \ln g(\mathbf{y}) - E_{g(\mathbf{y})} \ln p(\mathbf{y}|\boldsymbol{\theta}), \quad (1)$$

where $E_{g(\mathbf{y})}$ is the expectation with respect to $g(\mathbf{y})$.

Let $\boldsymbol{\theta}_n^p \in \boldsymbol{\Theta} \subset R^p$ be the pseudo true value that minimizes the KL loss between $g(\mathbf{y})$ and $p(\mathbf{y}|\boldsymbol{\theta})$

$$\boldsymbol{\theta}_n^p = \arg\min_{\boldsymbol{\theta}} KL[g(\mathbf{y}), p(\mathbf{y}|\boldsymbol{\theta})] = \arg\max_{\boldsymbol{\theta}} E_{g(\mathbf{y})} \ln p(\mathbf{y}|\boldsymbol{\theta}). \quad (2)$$

If the model is correctly specified, $\boldsymbol{\theta}_n^p$ is the true value, denoted $\boldsymbol{\theta}_0$.

Let $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ denote the quasi ML (QML) estimator of $\boldsymbol{\theta}_n^p$,[1] which maximizes the log-likelihood function of the parametric model,[2]

$$\widehat{\boldsymbol{\theta}}_n = \arg\max_{\theta} \ln p(\mathbf{y}|\boldsymbol{\theta}). \tag{3}$$

Let $\mathbf{y}^t = (y_1, y_2, \cdots, y_t)'$ denote all observed data at time t, $s_t(\boldsymbol{\theta}) = \partial \ln p(\mathbf{y}^t|\boldsymbol{\theta})/\partial\boldsymbol{\theta} - \partial \ln p(\mathbf{y}^{t-1}|\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ denote the t-th $P \times 1$ single point score vector at $\boldsymbol{\theta}$, $h_t(\boldsymbol{\theta}) = \partial s_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}'$ denote the t-th $P \times P$ single point Hessian matrix at $\boldsymbol{\theta}$. Define the sample Jacobian matrix at $\boldsymbol{\theta}$ as $\overline{\mathbf{J}}_n(\boldsymbol{\theta}) = \frac{1}{n}\sum_{t=1}^n s_t(\boldsymbol{\theta}) s_t(\boldsymbol{\theta})' - \frac{1}{n}\sum_{t=1}^n s_t(\boldsymbol{\theta}) \frac{1}{n}\sum_{t=1}^n s_t(\boldsymbol{\theta})'$, and define the sample Hessian-form information matrix at $\boldsymbol{\theta}$ as $\overline{\mathbf{I}}_n(\boldsymbol{\theta}) = -\overline{\mathbf{H}}_n(\boldsymbol{\theta}) = -\frac{1}{n}\sum_{t=1}^n h_t(\boldsymbol{\theta})$. White (1982) established the ML theory for misspecified models, that is,

$$\sqrt{n}\left[\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{J}}_n(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\right]^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p) \xrightarrow{d} N(\mathbf{0}, I_P), \tag{4}$$

where I_P stands for a P-dimensional identity matrix. So the asymptotic variance takes the sandwich form. If the model is correctly specified, then

$$\sqrt{n}[\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)]^{-1/2}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N(\mathbf{0}, I_P). \tag{5}$$

2.2. *Bayesian inference under model misspecification*

To do Bayesian inference about $\boldsymbol{\theta}$, let $p(\boldsymbol{\theta})$ be the prior distribution of $\boldsymbol{\theta}$. By Bayes' theorem, the standard posterior distribution is

$$p(\boldsymbol{\theta}|\mathbf{y}) = \frac{p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})}{p(\mathbf{y})} \propto p(\boldsymbol{\theta})p(\mathbf{y}|\boldsymbol{\theta}), \tag{6}$$

where $p(\mathbf{y}) = \int p(\mathbf{y}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}$ is the marginal likelihood. Unlike the ML theory, there is no difference between the Bayesian asymptotic theory for the correctly specified model and that for a misspecified model. In both cases, the Bayesian large sample theory (for example, Van der Vaart (1998)) guarantees that the scaled posterior distribution converges

[1] Given different parametric models, one may obtain distinct QML estimators. When focusing on a specific parametric model, the QML estimator is referred to as the ML estimator for the sake of simplicity.

[2] When there is no confusion, we simply write $\widehat{\boldsymbol{\theta}}_n(\mathbf{y})$ as $\widehat{\boldsymbol{\theta}}_n$.

to a normal distribution in total variation, that is,

$$\left\| p(\boldsymbol{\theta}|\mathbf{y}) - N\left(\widehat{\boldsymbol{\theta}}_n, \overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)/n\right)\right\|_{TV} \xrightarrow{p} 0, \tag{7}$$

where $\|p-q\|_{TV} = \int |p(x)-q(x)|dx$. So the posterior density can be approximated by a density of a Gaussian variate:

$$p^a(\boldsymbol{\theta}|\mathbf{y}) = \phi_{\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)/n}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n). \tag{8}$$

Observing the difference between the asymptotic posterior variance in (8) and the sandwich form of the asymptotic variance in (4), Müller (2013) proposed to make Bayesian inference based on an artificial posterior distribution, which is a Gaussian distribution centered at the MLE with the sandwich variance, i.e.,

$$p^s(\boldsymbol{\theta}|\mathbf{y}) = \phi_{\hat{\boldsymbol{\Sigma}}_s/n}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n), \tag{9}$$

where $\hat{\boldsymbol{\Sigma}}_s = \overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{J}}_n(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$.

Given a set of decisions $\mathcal{D}$ and a loss function $\mathcal{L}(\boldsymbol{\theta}, d)$ for any decision $d \in \mathcal{D}$, the optimal decision based on each of two posterior distributions is to minimize the posterior loss, i.e.,

$$d_a^*(\widehat{\boldsymbol{\theta}}_n) := \arg\min_{d\in\mathcal{D}} \int \mathcal{L}(\boldsymbol{\theta}, d)p^a(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta},\ d_s^*(\widehat{\boldsymbol{\theta}}_n) := \arg\min_{d\in\mathcal{D}} \int \mathcal{L}(\boldsymbol{\theta}, d)p^s(\boldsymbol{\theta}|\mathbf{y})\mathrm{d}\boldsymbol{\theta}.$$

The frequentist risk of the two optimal decisions may be obtained as

$$r\left(\boldsymbol{\theta}, d_a^*\right) = \int \mathcal{L}\left[\boldsymbol{\theta}, d_a^*(\widehat{\boldsymbol{\theta}}_n)\right] \phi_{\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)/n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\mathrm{d}\widehat{\boldsymbol{\theta}}_n,$$
$$r\left(\boldsymbol{\theta}, d_s^*\right) = \int \mathcal{L}\left[\boldsymbol{\theta}, d_s^*(\widehat{\boldsymbol{\theta}}_n)\right] \phi_{\hat{\boldsymbol{\Sigma}}_s/n}(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta})\mathrm{d}\widehat{\boldsymbol{\theta}}_n.$$

Müller (2013) showed that as $n \to +\infty$,

$$r\left(\boldsymbol{\theta}_n^p, d_s^*\right) \le r\left(\boldsymbol{\theta}_n^p, d_a^*\right).$$

Moreover, he showed that the inequality becomes strict for many loss functions. These findings imply that the Bayesian inference about the parameter of interest based on the standard posterior can be improved by that based on the sandwich posterior. A natural question to ask is whether the improvement also applies to model comparison from a Bayesian perspective. We first hope to answer this question.

3. RISK OF PREDICTIVE DISTRIBUTIONS ON MISSPECIFIED MODELS

3.1. *Predictive Distributions in Misspecified Models and the risk functions*

Given a parametric model $p(\mathbf{y}|\boldsymbol{\theta})$, which is potentially misspecified, we have different ways to predict future data, denoted by $\mathbf{y}_f$ whose density is $g(\mathbf{y}_f)$. For any predictive distribution whose density is $q(\mathbf{y}_f|\mathbf{y})$, following Aitchison (1975) and much of the literature, we use the KL loss function between $g\left(\mathbf{y}_f\right)$ and $q(\mathbf{y}_f|\mathbf{y})$, denoted by

$$KL\left[g\left(\mathbf{y}_f\right), q(\mathbf{y}_f|\mathbf{y})\right] = \int \ln \frac{g\left(\mathbf{y}_f\right)}{q(\mathbf{y}_f|\mathbf{y})} g\left(\mathbf{y}_f\right) \mathrm{d}\mathbf{y}_f, \tag{10}$$

to measure its predictive performance. It is important to find a good candidate model as well as a good predictive distribution that leads the smallest possible value to the KL loss. In this subsection, we introduce three different predictive distributions.[3]

The first predictive distribution is the plug-in predictive distribution, defined by

$$q(\mathbf{y}_f|\mathbf{y}) = p\left(\mathbf{y}_f|\bar{\boldsymbol{\theta}}_n\right), \tag{11}$$

where $\bar{\boldsymbol{\theta}}_n = \int \boldsymbol{\theta} p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}$ is the posterior mean.

The second predictive distribution is the Bayesian predictive distribution. It takes average on the parameter to eliminate the parameter uncertainty

$$q(\mathbf{y}_f|\mathbf{y}) = p(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\boldsymbol{\theta}, \mathbf{y}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}. \tag{12}$$

Here the 'average' is taken on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$ that does not take account of the model misspecification.

Thirdly, if we replace the posterior distribution by the sandwich posterior distribution of Müller (2013), then we get a new predictive distribution

$$q(\mathbf{y}_f|\mathbf{y}) = p^s(\mathbf{y}_f|\mathbf{y}) = \int p(\mathbf{y}_f|\boldsymbol{\theta}, \mathbf{y}) p^s(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta}, \tag{13}$$

[3] The predictive density estimation and the comparison framework under the KL loss was initially established in Aitchison and Dunsmore (1975) and Aitchison (1975). The framework has been applied in many fields, including decision theory, information theory, econometrics, machine learning, image processing, and mathematical finance. Notable contributions include, but are not limited to, Komaki (2001), George et al. (2006), Brown et al. (2008), Kato (2009), Marchand and Sadeghkhani (2018), Hamura and Kubokawa (2022) and Nishi et al. (2024).

which is termed as the 'sandwich Bayesian predictive distribution'. To the best of our knowledge, this predictive distribution has not yet been used in the predictive literature.

Following the existing literature, we assume that $\mathbf{y}_f$ is independent of $\mathbf{y}$, which we denote by $\mathbf{y}_{rep}$. Given three different predictive distributions, a natural question arises: for a misspecified model, which of the three predictive distributions one should use to get the best prediction?

In this paper, we compare alternative predictive distributions from a decision-theoretical viewpoint. Let $\mathcal{D} = \{d_a\}_{a=1}^{3}$ be the set of decisions, where d_a is the decision of using (11) or (12) or (13) if $a = 1$ or 2 or 3. The predictive distribution under different action can be expressed as $p(\mathbf{y}_{rep}|\mathbf{y}, d_a)$.

For candidate model M that is potentially misspecified, let $p(\mathbf{y}_{rep}|\mathbf{y},M,d_a)$ be the predictive distribution under decision d_a and model M, and let the loss function $\mathcal{L}(\mathbf{y},M,d_a)$ be 2 times the KL loss function between $g(\mathbf{y}_f)$ and $p(\mathbf{y}_{rep}|\mathbf{y},M,d_a)$, that is,

$$\mathcal{L}(\mathbf{y},M,d_a) = 2 \times KL\left[g(\mathbf{y}_f), p(\mathbf{y}_{rep}|\mathbf{y},M,d_a)\right] = 2\int \ln \frac{g(\mathbf{y}_{rep})}{p(\mathbf{y}_{rep}|\mathbf{y},M,d_a)} g(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}_{rep}.$$

Given the loss function, the frequentist (average) risk of decision d_a under model M is (Good (1952) and Aitchison (1975))[4]

$$Risk(M,d_a) = E_{g(\mathbf{y})}\left[\mathcal{L}(\mathbf{y},M,d_a)\right] = \int g(\mathbf{y}) \int \ln \frac{g(\mathbf{y}_{rep})}{p(\mathbf{y}_{rep}|\mathbf{y},M,d_a)} g(\mathbf{y}_{rep})\mathrm{d}\mathbf{y}_{rep}\mathrm{d}\mathbf{y}.$$

Hence, the selection of a predictive distribution and a candidate model becomes

$$(a^*, M^*) = \arg\min_{a,M} \left\{2E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})} \ln g(\mathbf{y}_{rep}) - 2E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})} \ln p(\mathbf{y}_{rep}|\mathbf{y}, M, d_a)\right\}.$$

Since $g(\mathbf{y}_{rep})$ is the DGP and $E_{g(\mathbf{y}_{rep})} \ln g(\mathbf{y}_{rep})$ is independent of candidate models and predictive distributions, the selection problem is the same as

$$(a^*, M^*) = \arg\min_{a,M} \left\{-2E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})} \ln p(\mathbf{y}_{rep}|\mathbf{y}, M, d_a)\right\}. \tag{14}$$

and the frequentist risk of decision d_a under model M can be equivalently written as

$$Risk(M,d_a) = -2E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})} \ln p(\mathbf{y}_{rep}|\mathbf{y}, M, d_a). \tag{15}$$

[4] When there is no confusion, we purge M from $Risk(M,d_a)$.

In this case, the predictive distribution corresponding to statistical decision d_{a^*} and the optimal model M^* is selected. The smaller $Risk(M, d_a)$, the better the predictive distribution performs when using $p(\mathbf{y}_{rep}|\mathbf{y}, M, d_a)$ to predict $g(\mathbf{y}_{rep})$.[5]

3.2. *Risks of Three Different Predictive Distributions in Misspecified Model*

In this subsection, we give the regularity conditions and make the frequent risk analysis for these three predictive distributions. Before we introduce regular conditions, we need to fix some notations. Let $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right) = \ln p(\mathbf{y}^t|\boldsymbol{\theta}) - \ln p(\mathbf{y}^{t-1}|\boldsymbol{\theta})$ be the conditional log-likelihood for the t-th observation for any $1 \leq t \leq n$. For simplicity, we suppress $l_t\left(\mathbf{y}^t, \boldsymbol{\theta}\right)$ as $l_t(\boldsymbol{\theta})$ so that the log-likelihood function $\ln p(\mathbf{y}|\boldsymbol{\theta})$ is $\sum_{t=1}^{n} l_t(\boldsymbol{\theta})$.[6] Define $l_t^{(j)}(\boldsymbol{\theta})$ to be the j-th derivative of $l_t(\boldsymbol{\theta})$ and

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

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

$$\mathbf{B}_n(\boldsymbol{\theta}) := \mathrm{Var}\left[\frac{1}{\sqrt{n}} \sum_{t=1}^{n} l_t^{(1)}(\boldsymbol{\theta})\right], \mathbf{H}_n(\boldsymbol{\theta}) := \int \overline{\mathbf{H}}_n(\boldsymbol{\theta}) g(\mathbf{y}) \, d\mathbf{y}, \ \mathbf{J}_n(\boldsymbol{\theta}) = \int \overline{\mathbf{J}}_n(\boldsymbol{\theta}) g(\mathbf{y}) \, d\mathbf{y}.$$

The regularity conditions we impose are similar to those in Li et al. (2020). For the detailed discussion of these conditions, see Li et al. (2020).

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

Assumption 2: $\{y_t\}_{t=1}^{\infty}$ is strong mixing 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, $l_t(\boldsymbol{\theta})$ is third-times differentiable on $\boldsymbol{\Theta}$ almost surely.

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

[5] It should be noted that the predictive method d_a and the candidate model M are jointly optimized in (14).

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

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

Assumption 6: $\left\{ l_t^j(\boldsymbol{\theta}) \right\}$ is L_2-near epoch dependent with respect to $\{\mathbf{y}_t\}$ of size -1 for $0 \leq j \leq 1$ and $-\frac{1}{2}$ for $j = 2, 3$ uniformly on $\boldsymbol{\Theta}$.

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

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

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

$$\lim_{n\to\infty} \sup \sup_{\Theta \setminus N(\boldsymbol{\theta}_n^p, \varepsilon)} \frac{1}{n} \sum_{t=1}^{n} \left\{ E\left[l_t(\boldsymbol{\theta}) \right] - E\left[l_t(\boldsymbol{\theta}_n^p) \right] \right\} < 0, \tag{16}$$

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 $0 < p\left(\boldsymbol{\theta}_n^0\right) < \infty$ 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 $\int \|\boldsymbol{\theta}\|^2 p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} < \infty$.

For simplification, we denote $\mathbf{H}_n(\boldsymbol{\theta}_n^p)$ as $\mathbf{H}_n$, $\mathbf{B}_n(\boldsymbol{\theta}_n^p)$ as $\mathbf{B}_n$. Given these regularity conditions, we can derive the asymptotic approximation for risks associated with the three predictive distributions, which are given by[8]

Plug-in prediction: $$Risk(d_1) = E_{g(\mathbf{y})} E_{g(\mathbf{y}_{rep})} \left[-2 \ln p(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}) \right],$$

Bayesian prediction: $$Risk(d_2) = E_{g(\mathbf{y})} E_{g(\mathbf{y}_{rep})} \left[-2 \ln \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right],$$

Sandwich Bayesian prediction: $$Risk(d_3) = E_{g(\mathbf{y})} E_{g(\mathbf{y}_{rep})} \left[-2 \ln \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}) p^s(\boldsymbol{\theta}|\mathbf{y}) d\boldsymbol{\theta} \right].$$

[7] Here we denote the pseudo-true value as $\boldsymbol{\theta}_n^p$ to allow it varies with the sample size n. This notation can accommodate to the dependence and heterogeneity of data.

[8] To simplify notations, we purge their dependence on candidate model M.

Note that the sandwich posterior is given by

$$p^s(\boldsymbol{\theta}|\mathbf{y}) = \phi_{\widehat{\boldsymbol{\Sigma}}_n/n}(\boldsymbol{\theta} - \widehat{\boldsymbol{\theta}}_n),\ \widehat{\boldsymbol{\Sigma}}_n = \widehat{\mathbf{H}}_n^{-1}\widehat{\mathbf{B}}_n\widehat{\mathbf{H}}_n^{-1},$$

where $\widehat{\mathbf{H}}_n$ and $\widehat{\mathbf{B}}_n$ are consistent estimators of $\mathbf{H}_n$ and $\mathbf{B}_n$. An example is, taking $\widehat{\mathbf{H}}_n = -\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$ and $\widehat{\mathbf{B}}_n = \overline{\mathbf{J}}_n(\widehat{\boldsymbol{\theta}}_n)$ in the independent case, $\widehat{\boldsymbol{\Sigma}}_n = \overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{J}}_n(\widehat{\boldsymbol{\theta}}_n)\overline{\mathbf{I}}_n^{-1}(\widehat{\boldsymbol{\theta}}_n)$. In the later sections, different choices will be discussed. Here we write $\widehat{\mathbf{H}}_n$ and $\widehat{\mathbf{B}}_n$ for generality.

In this subsection, we will give the asymptotic approximation for $Risk(d_1), Risk(d_2)$ and $Risk(d_3)$. Note that the risk of the plug-in predictive distribution, $Risk(d_1)$, has been used in Li et al. (2020) to develop the new DIC criteria for comparing misspecified models. They derived the asymptotic approximation for $Risk(d_1)$ under similar regularity conditions. That is

$$Risk(d_1) = -2E_{g(\mathbf{y})}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] + 2\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right] + o\left(1\right). \tag{17}$$

Hence, an asymptotically unbiased estimator of $Risk(d_1)$ is

$$-2\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right) + 2\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right].$$

In the following two theorems, we derive asymptotically unbiased estimators for $Risk(d_2)$ and $Risk(d_3)$, which are our core theoretical results.

THEOREM 1: *Under Assumptions 1-9, it can be shown that*

$$Risk(d_2) = -2E_{g(\mathbf{y})}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] + \mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right] + P\ln 2 + o\left(1\right). \tag{18}$$

For i.i.d. data, Ando and Tsay (2010) gave an alternative expression as

$$-2E_{g(\mathbf{y})}\left[\ln\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}\right] + \mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]. \tag{19}$$

Note that $\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}$ is the Bayesian predictive distribution of $\mathbf{y}$ taking average on the posterior distribution $p(\boldsymbol{\theta}|\mathbf{y})$. It can be shown that

$$E_{g(\mathbf{y})}\left[\ln\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}\right] = E_{g(\mathbf{y})}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] - \frac{1}{2}P\ln 2 + o\left(1\right). \tag{20}$$

Thus, these two expressions are asymptotically equivalent. From (20), the first term of (19), $E_{g(\mathbf{y})}\left[\ln\int p\left(\mathbf{y}|\boldsymbol{\theta}\right)p\left(\boldsymbol{\theta}|\mathbf{y}\right)d\boldsymbol{\theta}\right]$, cannot be interpreted as a model fit term because it includes

$P \ln 2$ which is a penalty term. So (18) bears more similarity to the traditional information criteria, such as AIC, TIC and DIC.

THEOREM 2: *Under Assumptions 1-9, it can be shown that*

$$Risk(d_3) = -2E_{g(\mathbf{y})}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] + \mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right] + \ln\left(\left|\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1} + I_P\right|\right) - \mathbf{tr}\left[\left(\mathbf{B}_n + \mathbf{H}_n\right)\left(-\mathbf{H}_n + \boldsymbol{\Sigma}_n^{-1}\right)^{-1}\right] + o(1), \quad (21)$$

where $\boldsymbol{\Sigma}_n = \mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}$.

REMARK 1: From Theorem 1 and Theorem 2, we can get asymptotically unbiased estimators for $Risk(d_2)$ and $Risk(d_3)$ by taking sample analogs respectively. These two asymptotically unbiased estimators provide the basis for developing information criteria for Bayesian model comparison, which will be reported in next section.

Note that the plug-in predictive distribution deals with model misspecification by plugging in the posterior mean, which is asymptotically normal distributed with the variance being a sandwich covariance matrix. The Bayesian predictive distribution is

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

It takes account of the influence of parameter uncertainty via the standard posterior distribution $p\left(\boldsymbol{\theta}|\ \mathbf{y}\right)$ and handle model misspecification via $p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)$. Furthermore, the sandwich Bayesian predictive distribution is

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

It takes account of the influence of parameter uncertainty via the sandwich posterior distribution $p^s\left(\boldsymbol{\theta}|\mathbf{y}\right)$ and handle model misspecification with both $p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)$ and the sandwich posterior distribution $p^s\left(\boldsymbol{\theta}|\mathbf{y}\right)$.

Clearly, the three estimators for the risk functions shares the first term, $-2\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)$, which measure the model fit. They also share the second term, $\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]$ which measures model misspecification. The third term in the estimator of $Risk(d_2)$ is $P\ln 2$ that measures the influence of the parameter uncertainty based on the standard posterior distribution. The third term in the estimator of $Risk(d_3)$ is $\ln\left(\left|\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1} + I_P\right|\right) -$

$\mathbf{tr}\left[\left(\mathbf{B}_n+\mathbf{H}_n\right)\left(-\mathbf{H}_n+\mathbf{\Sigma}_n^{-1}\right)^{-1}\right]$ that measures the influence of the parameter uncertainty based on the sandwich posterior distribution with the consideration of the model misspecification.

REMARK 2: The plug-in predictive distribution does not consider the parameter uncertainty, while both the Bayesian predictive distribution and the sandwich Bayesian predictive distribution take parameter uncertainty into account by taking an average. The only difference between the last two distributions is that the Bayesian predictive distribution takes the average over the standard posterior distribution, while the sandwich Bayesian predictive distribution takes average over the sandwich Bayesian posterior, which has been adjusted for model misspecification.

COROLLARY 3: Under Assumptions 1-9, when the model is misspecified such that

$$\lim_{n\to\infty}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]\geq P, \tag{22}$$

then

$$\lim_{n\to\infty}\left(Risk(d_2)-Risk(d_1)\right)<0.$$

REMARK 3: This corollary gives a sufficient condition under which the Bayesian predictive distribution is better than the plug-in predictive distribution asymptotically. In fact, we can rewrite (22) as

$$\lim_{n\to\infty}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]-P=\lim_{n\to\infty}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}-I_P\right]=\lim_{n\to\infty}\mathbf{tr}\left[\left(\mathbf{B}_n+\mathbf{H}_n\right)\left(-\mathbf{H}_n\right)^{-1}\right]\geq 0.$$

Thus, a sufficient condition to ensure (22) is that $\mathbf{B}_n+\mathbf{H}_n$ is positive definite uniformly in n.[9]

[9]Let $P\times P$ matrices A and B be symmetric and positive definite. Hence, there exists a $P\times P$ matrix Q such that $B=QQ^T$, and

$$\mathbf{tr}\left(AB\right)=\mathbf{tr}\left(AQQ^T\right)=\mathbf{tr}\left(QAQ^T\right)=\sum_{j=1}^{P}q_j'Aq_j>0,$$

where q_j is the j-th column vector of Q.

REMARK 4: When the model is correctly specified, we have $\lim_{n\to\infty} \mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right] = P$ and hence,

$$Risk(d_1) = -2E_{g(\mathbf{y})}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] + 2P + o\left(1\right), \tag{23}$$

$$Risk(d_2) = -2E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right] + P(1+\ln 2) + o\left(1\right). \tag{24}$$

Since $P\ln 2 < P$, we have

$$\lim_{n\to\infty}\left(Risk(d_2) - Risk(d_1)\right) < 0.$$

This suggests the predictive distribution $p\left(\mathbf{y}_{rep}|\mathbf{y}\right)$ has a lower asymptotic risk than the plug-in predictive distribution. The limitation of the plug-in predictive distributions stems from their failure to account for parameter uncertainty, as they treat parameters as the estimated quantities. In contrast, the Bayesian method embrace this uncertainty by integrating out parameters with respect to their posterior distribution. For detailed discussions of the parameter uncertainty, readers may refer to Barberis (2000) and George and Xu (2010). It is easy to show that replacing $\overline{\boldsymbol{\theta}}_n$ with $\widehat{\boldsymbol{\theta}}_n$ in (23) and (24) does not change the results.

Under the assumption of correct model specification, the comparison of the Bayesian predictive distributions and the plug-in alternatives in terms of the frequentist risk has been extensively studied in the statistics literature. Most of them focus their attention to specific model setups or to specific prior distributions.

For example, in finite samples, Aitchison (1975) showed that the MLE-plug-in predictive distribution for Gamma and normal models are uniformly dominated by Bayesian predictive distribution with uniform priors. Murray (1977) and Ng (1980) showed that the Bayesian predictive density with uniform priors is the best predictive distribution that is invariant under the translation group. Levy and Perng (1986) proved that the Bayesian predictive distribution with a diffuse prior dominates the plug-in predictive distribution for normal linear models.

From an asymptotic point of view, Komaki (1996) showed that, for the multidimensional curved exponential family, the plug-in predictive distribution with the asymptotically efficient estimators can generate the frequentist risk that asymptotically coincide with that of the Bayesian predictive distributions. For multivariate normal models with unknown

means, Komaki (2001) proved that the Bayesian predictive distribution with Stein's prior dominates both the Bayesian predictive distribution with a uniform prior and the plug-in predictive distribution. George et al. (2006) showed that any Bayes predictive density is minimax if it is obtained by a prior yielding a marginal that is superharmonic or whose square root is superharmonic for multivariate normal models with unknown means. For multivariate normal linear models, George and Xu (2008) obtained sufficient conditions of the Bayesian predictive distribution with different priors for minimaxity and dominance over the Bayes predictive distribution with uniform priors and the plug-in predictive distribution. For multivariate models with unknown means and variances, Kato (2009) proposed to use an improper shrinkage prior with which the Bayesian predictive distribution dominates the Bayes predictive distribution with uniform priors and the plug-in predictive distribution. For multivariate normal models with unknown means whose parameter space restricted to a convex set, Fourdrinier et al. (2011) showed that the Bayesian predictive distribution with a uniform prior on the convex set dominates the plug-in predictive distribution. Matsuda and Komaki (2015) developed singular value shrinkage priors for the mean matrix parameters in the matrix variate normal model with known covariance matrices and showed that the Bayesian predictive distributions based on these priors are minimax and dominate those based on uniform priors and the plug-in predictive distributions. For multivariate normal models with additional information for means and variances, Marchand and Sadeghkhani (2018) gave the conditions under which the Bayesian predictive distribution with uniform prior defined on the information set dominates the plug-in predictive distribution. For Type-II censored data that is generated by ordered observations, Nishi et al. (2024) prove that the Bayesian predictive distribution with an improper Gamma prior dominates the plug-in predictive distribution.

Almost all of these works are about normal models or normal linear models, but our work give an asymptotic results for much general class of models. Moreover, none of these studies allow model misspecification. When the model is misspecified, the claim of dominance of the Bayesian predictive distribution using the standard posterior over the plug-in predictive distribution may not valid.

REMARK 5: When the model is misspecified, we argue that in Corollary 3 the condition $\lim_{n\to\infty} \mathbf{tr}\left[\mathbf{B}_n \left(-\mathbf{H}_n\right)^{-1}\right] \geq P$ can be satisfied in most cases. To see this, note that the

asymptotic covariance matrix of QMLE is $\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}$, while the asymptotic covariance matrix of MLE is $-\mathbf{H}_n^{-1}$. QMLE is more robust than MLE, while the price we pay is the loss of efficiency. Thus, we expect that $\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1} - (-\mathbf{H}_n^{-1}) = \mathbf{H}_n^{-1}(\mathbf{B}_n + \mathbf{H}_n)\mathbf{H}_n^{-1} \geq 0$, that is $\mathbf{B}_n + \mathbf{H}_n \geq 0$ and $\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n^{-1})\right] \geq P$. This argument is based on the empirical phenomenon that "robust standard error is often larger than simple standard error" which is often observed in empirical research; for instance, White's robust standard errors and the cluster-robust standard errors are typically larger than the simple OLS standard errors in most cases (Angrist and Pischke (2009)).

COROLLARY 4: Under Assumptions 1-9, it can be shown that

$$\lim_{n\to\infty}\left(Risk(d_3) - Risk(d_2)\right) \leq 0.$$

REMARK 6: Corollary 4 shows that when the model is misspecified, the risk of Müller's sandwich predictive posterior distribution is always less (weakly) than that of the original Bayesian predictive distribution in terms of KL loss function asymptotically. This can explained by the arguments in Remark 2. The original Bayesian predictive distribution only considers the parameter uncertainty, while the sandwich Bayesian predictive distribution considers both parameter uncertainty and model uncertainty by replacing the posterior $p(\boldsymbol{\theta}|\mathbf{y})$ by the sandwich posterior $p^s(\boldsymbol{\theta}|\mathbf{y})$. The result in Corollary 4 extends the result of Müller (2013) to model comparison.

COROLLARY 5: Under Assumptions 1-9, when the model is misspecified such that

$$\lim_{n\to\infty}\mathbf{tr}\left[\left(\mathbf{B}_n + \mathbf{H}_n\right)\left(-\mathbf{H}_n + \boldsymbol{\Sigma}_n^{-1}\right)^{-1}\right] \geq 0, \tag{25}$$

it can be shown that

$$\lim_{n\to\infty}\left(Risk(d_3) - Risk(d_1)\right) \leq 0.$$

REMARK 7: This theorem gives sufficient conditions under which the sandwich Bayesian predictive distribution can achieve a lower risk than the plug-in predictive distribution asymptotically. In fact, a sufficient condition for (25) is that $\mathbf{B}_n + \mathbf{H}_n$ is positive definite uniformly in n, which is consistent with the trace condition in Theorem 3.

REMARK 8: Under Assumptions 1-9, when the model is correctly specified, the information equality holds. Consequently,,

$$\lim_{n\to\infty}(Risk(d_2) - Risk(d_3)) = 0,$$
$$\lim_{n\to\infty}(Risk(d_3) - Risk(d_1)) < 0.$$

This suggests that, when model is correctly specified, the sandwich Bayesian predictive distribution is asymptotic equivalent with the Bayesian predictive distribution, while both of them are better than the plug-in predictive distribution.

3.3. *A Toy Model*

To illustrate the risks associated with the three predictive distribution, we consider the following toy model. The data $\{y_i\}_{i=1}^n$ are observed and the k-dimensional explanatory variable $\{X_i\}_{i=1}^n$ are fixed for simplification. Suppose the true DGP is a linear model with heteroskedasticity:

$$y_i = X_i'\beta + \varepsilon_i, \varepsilon_i \sim N(0, \sigma_i^2).$$

The k-dimension regression coefficient β is of interest. However, since we do not know the true DPG, we assume the following misspecified linear regression model with homoskedasticity is used to fit the data:

$$y_i = X_i'\beta + \varepsilon_i, \varepsilon_i \sim N(0, \sigma^2). \tag{26}$$

The variance $\sigma^2 = \sum_{i=1}^n \sigma_i^2/n$ is assumed to be known.[10] Let $\mathbf{y} = (y_1, ..., y_n)'$, $\mathbf{X} = (X_1, X_2, ..., X_n)'$, then the log-likelihood function is given as

$$\ln p(\mathbf{y}|\mathbf{X}, \beta) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - X_i'\beta)^2.$$

The QMLE of β is given by $\hat{\beta} = \left(\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\sum_{i=1}^n X_i y_i\right)$, which is also the ordinary least square (OLS) estimator and the posterior mean under standard priors.

[10] If σ^2 is unknown, it can be shown that $(\beta, \sum_{i=1}^n \sigma_i^2/n)$ is the pseudo-true value. This simplification will not affect the key conclusion.

Denote $Q_n = \left(\sum_{i=1}^n X_i X_i'\right)/n$ and $V_n = \left(\sum_{i=1}^n \sigma_i^2 X_i X_i'\right)/n$. We then have

$$\mathbf{B}_n = Var\left(n^{-1/2}\frac{d\ln p(\mathbf{y}|\mathbf{X},\beta)}{d\beta}\right) = Var\left(\frac{1}{\sqrt{n}\sigma^2}\sum_{i=1}^n X_i\varepsilon_i\right) = V_n/\sigma^4,$$

$$\mathbf{H}_n = E\left(\frac{1}{n}\frac{d^2\ln p(\mathbf{y}|\mathbf{X},\beta)}{d\beta d\beta^{'}}\right) = -Q_n/\sigma^2.$$

In our model (26), the variance structure is misspecified. If the heteroskedasticity is absent, i.e., if $\sigma_1^2 = ... = \sigma_n^2 = \sigma^2$, then $V_n = \left(\sum_{i=1}^n \sigma_i^2 X_i X_i'\right)/n = \sigma^2 Q_n$. In this case, the information equality $\mathbf{B}_n + \mathbf{H}_n = 0$ holds. However, heteroskedasticity breaks the information equality, that is,

$$\mathbf{B}_n + \mathbf{H}_n = \frac{1}{\sigma^4}\left[\frac{1}{n}\sum_{i=1}^n \sigma_i^2 X_i X_i' - \left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2\right)\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)\right] \neq 0.$$

When the condition $\mathbf{tr}[\mathbf{B}_n(-\mathbf{H}_n)^{-1}] \geq k$ holds, note that the covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$ is

$$\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1} = \left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^n \sigma_i^2 X_i X_i'\right)\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1},$$

which is White's heteroskedasticity robust covariance matrix for QMLE in White (1980). Note that the covariance matrix of $\sqrt{n}(\hat{\beta} - \beta)$, ignoring heteroskedasticity, is

$$-\mathbf{H}_n^{-1} = \sigma^2\left(\frac{1}{n}\sum_{i=1}^n X_i X_i'\right)^{-1},$$

which should be smaller than White's heteroskedasticity robust covariance matrix.[11] That is,

$$\mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1} \geq -\mathbf{H}_n^{-1},$$

that is $\mathbf{B}_n + \mathbf{H}_n \geq 0$. So we expect the trace condition $\mathbf{tr}[\mathbf{B}_n(-\mathbf{H}_n)^{-1}] \geq k$ holds and Corollary 3 can be applied. Hence, the Bayesian predictive distribution has smaller risk than the plug-in predictive distribution. What is more, Corollary 4 guarantees the sandwich

[11] A sufficient condition is that $\sigma_i^2 = \sigma^2(X_i)$ and $X_i X_i^{'}$ are positively correlated.

Bayesian predictive distribution has a smaller risk than the Bayesian predictive distribution. That is,

$$Risk(d_3) \leq Risk(d_2) \leq Risk(d_1).$$

for sufficient large n.

In this toy model, we can verify this result hold exactly for every n, because we can directly derive the risk of all three predictive distributions and compare them. Consider the independent replication data:

$$y_{rep,i} = X_i'\beta + \varepsilon_{rep,i},\ \varepsilon_{rep,i} \sim N(0, \sigma_i^2),\ i = 1, 2, ..., n.$$

where $\varepsilon_{rep,1}, ..., \varepsilon_{rep,n}$ is independent of $\varepsilon_1, ..., \varepsilon_n$.

Let $\mathbf{y}_{rep} = (y_{rep,1}, ..., y_{rep,n})'$, the plug-in predictive distribution is

$$p(\mathbf{y}_{rep}|\mathbf{X}, \hat{\beta}) = N(\hat{\beta}, \sigma^2 I_n), \hat{\beta} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i y_i.$$

Then we get the risk of the loss associated with the plug-in predictive distribution:

$$\begin{aligned} Risk(d_1) &= E_{g(\mathbf{y})} E_{g(\mathbf{y}_{rep})} \left[-2 \ln p(\mathbf{y}_{rep}|\mathbf{X}, \hat{\beta})\right] \\ &= E_{g(\mathbf{y})} E_{g(\mathbf{y}_{rep})} \left[n \ln(2\pi\sigma^2) + \frac{1}{\sigma^2} \sum_{i=1}^{n} (y_{rep,i} - X_i'\hat{\beta})\right] \\ &= n[\ln(2\pi\sigma^2) + 1] + \mathbf{tr}\left[\sigma^{-2} Q_n^{-1} V_n\right]. \end{aligned} \tag{27}$$

Note that $\mathbf{tr}\left[\sigma^{-2} Q_n^{-1} V_n\right] = \mathbf{tr}\left[\mathbf{B}_n \left(-\mathbf{H}_n\right)^{-1}\right]$, and that

$$E_{g(\mathbf{y})}\left[-2 \ln p(\mathbf{y}|\mathbf{X}, \hat{\beta})\right] = n[\ln(2\pi\sigma^2) + 1] - \mathbf{tr}\left[\sigma^{-2} Q_n^{-1} V_n\right].$$

The asymptotic expansion (17) holds, that is

$$Risk(d_1) = E_{g(\mathbf{y})}\left[-2 \ln p(\mathbf{y}|\mathbf{X}, \hat{\beta})\right] + 2\mathbf{tr}\left[\mathbf{B}_n \left(-\mathbf{H}_n\right)^{-1}\right].$$

For the Bayesian predictive distribution, we should calculate the posterior distribution $p(\beta|\mathbf{y}, \mathbf{X})$. For simplification, we use the flat prior $p(\beta) \propto 1$. Then the posterior distribution

is

$$\beta|\mathbf{y},\mathbf{X} \sim N(\hat{\beta}, \sigma^2 Q_n^{-1}/n).$$

So the Bayesian predictive distribution is given by

$$p(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X}) = \int p(\mathbf{y}_{rep}|\mathbf{X},\beta)p(\beta|\mathbf{y},\mathbf{X})d\beta.$$

Let

$$\hat{\beta}_{rep} = \left(\sum_{i=1}^{n} X_i X_i'\right)^{-1} \sum_{i=1}^{n} X_i y_{rep,i}.$$

It can be shown that

$$\ln p(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X}) = -\frac{n}{2}\ln(2\pi\sigma^2) + \frac{1}{2}\ln\left|\frac{1}{2}I_k\right| + \frac{1}{4\sigma^2}\left(\hat{\beta}_{rep}+\hat{\beta}\right)' nQ_n\left(\hat{\beta}_{rep}+\hat{\beta}\right) - \frac{1}{2\sigma^2}\left[\sum_{i=1}^{n}(y_{rep,i})^2 + \hat{\beta}' nQ_n\hat{\beta}\right].$$

Taking expectation with respect to $\mathbf{y}$ and $\mathbf{y}_{rep}$, we get the risk associated with the Bayesian predictive distribution:

$$Risk(d_2) = E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})}\left[-2\ln p(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X})\right] = n[\ln(2\pi\sigma^2)+1] + k\ln 2. \tag{28}$$

So the asymptotic expansion in Lemma 1 holds exactly, that is

$$Risk(d_2) = E_{g(\mathbf{y})}\left[-2\ln p(\mathbf{y}|\mathbf{X},\hat{\beta})\right] + \mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right] + k\ln 2.$$

Given the trace condition $\mathbf{tr}[\mathbf{B}_n(-\mathbf{H}_n)^{-1}] \geq k$, we have $Risk(d_2) \leq Risk(d_1)$ exactly holds in this example.

The sandwich Bayesian predictive distribution is given by

$$p^s(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X}) = \int p(\mathbf{y}_{rep}|\mathbf{X},\beta)p^s(\beta|\mathbf{y},\mathbf{X})d\beta,$$

where $p^s(\beta|\mathbf{y},X)$ is the density of $N(\hat{\beta}, Q_n^{-1}V_nQ_n^{-1}/n)$ evaluated at β. Hence,

$$\ln p^s(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X}) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2}\ln\left|\sigma^{-2}V_nQ_n^{-1}I_k\right|$$

$$+\frac{n}{2}\left(\sigma^{-2}\hat{\beta}_{rep}+V_n^{-1}Q_n\hat{\beta}\right)'\left(\sigma^{-2}Q_n^{-1}+V_n^{-1}\right)^{-1}\left(\sigma^{-2}\hat{\beta}_{rep}+V_n^{-1}Q_n\hat{\beta}\right)$$
$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_{rep,i})^2-\frac{n}{2}\hat{\beta}'Q_nV_n^{-1}Q_n\hat{\beta}.$$

Taking expectation with respect to $\mathbf{y}$ and $\mathbf{y}_{rep}$, we get

$$\begin{aligned}Risk(d_3) &= E_{g(\mathbf{y})}E_{g(\mathbf{y}_{rep})}\left[-2\ln p^s(\mathbf{y}_{rep}|\mathbf{y},\mathbf{X})\right]\\ &= n[\ln(2\pi\sigma^2)+1]+\ln\left|\sigma^{-2}V_nQ_n^{-1}+I_k\right|\\ &\quad-\mathbf{tr}\left[\sigma^2(\sigma^{-2}V_n-Q_n)(\sigma^{-2}Q_n+Q_nV_n^{-1}Q_n)^{-1}\right]. \qquad (29)\end{aligned}$$

So the asymptotic expansion in Lemma 2 holds exactly. It can be verified that $Risk(d_3) \leq Risk(d_2)$ because of $\mathbf{B}_n+\mathbf{H}_n \geq 0$. This is easy to understand, because the sandwich Bayesian predictive distribution is based on the sandwich posterior, which is adjusted for the model misspecification.

4. BAYESIAN PREDICTIVE INFORMATION CRITERIA FOR COMPARING MISSPECIFIED MODELS

4.1. *Statistical Decision Theory for Model Selection*

In this section, from the predictive viewpoint, we develop new information criteria for Bayesian model comparison. Suppose there are k candidate models $M_1, M_2, \cdots, M_K$ that are all potentially misspecified and we hope to select a model from the pool.

To begin, we define some notations. For P_k-dimension candidate model M_k, the vector of parameters is $\boldsymbol{\theta}_k \in \boldsymbol{\Theta}_k \subset R^{P_k}$ and $p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)$ is applied to fit the data. The posterior distribution of model M_k is denoted as $p(\boldsymbol{\theta}_k|\mathbf{y}, M_k)$, the pseudo-true value, QMLE, posterior mean of model M_k are denoted as $\boldsymbol{\theta}_n^k$, $\widehat{\boldsymbol{\theta}}_n^k$ and $\bar{\boldsymbol{\theta}}_n^k$, respectively. $\mathbf{B}_n^k, \mathbf{H}_n^k, \boldsymbol{\Sigma}_n^k, \widehat{\mathbf{B}}_n^k, \widehat{\mathbf{H}}_n^k, \widehat{\boldsymbol{\Sigma}}_n^k, p^s(\boldsymbol{\theta}_k|\mathbf{y}, M_k)$ can be defined in the same way.

The traditional model selection argument considers how to choose the 'best' model among them. However, we propose to choose the best model and the best predictive distribution, that is,

$$\min_{a\in\{1,2,3\},k\in\{1,\cdots,K\}} Risk(M_k, d_a) = E_{g(\mathbf{y})}\left(2\times KL\left[g(\mathbf{y}_{rep}), p(\mathbf{y}_{rep}|\mathbf{y}, M_k, d_a)\right]\right). \qquad (30)$$

Note that $p(\mathbf{y}_{rep}|\mathbf{y}, M_k, d_a)$ denotes the predictive distribution under predictive decision d_a and model M_k, $Risk(M_k, d_a)$ denotes the corresponding predictive risk. To be specific,

$$p(\mathbf{y}_{rep}|\mathbf{y}, M_k, d_1) = p\left(\mathbf{y}_{rep}|\overline{\boldsymbol{\theta}}_n^k, M_k\right) \tag{31}$$

$$p(\mathbf{y}_{rep}|\mathbf{y}, M_k, d_2) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}_k, M_k)\, p(\boldsymbol{\theta}_k|\mathbf{y}, M_k)\, d\boldsymbol{\theta}_k \tag{32}$$

$$p(\mathbf{y}_{rep}|\mathbf{y}, M_k, d_3) = \int p(\mathbf{y}_{rep}|\boldsymbol{\theta}_k, M_k)\, p^s(\boldsymbol{\theta}_k|\mathbf{y}, M_k)\, d\boldsymbol{\theta}_k \tag{33}$$

The optimization problem (30) simultaneously solve the best model and the best predictive distribution.

To the best of our knowledge, there are two information criteria that allow for model misspecification in the literature, TIC of Takeuchi (1976) and DIC_M of Li et al. (2020), both of which assume that the predictive distribution is the plug-in distribution. $\text{DIC}_M(k)$ takes the form of

$$\text{DIC}_M(k) = -2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) + 2P_M^k, \text{ with } P_M^k = \mathbf{tr}\left\{n\overline{\boldsymbol{\Omega}}_n^k \mathbf{V}_n^k\right\}, \tag{34}$$

where $n\mathbf{V}_n^k = nE\left[\left(\boldsymbol{\theta}_k - \overline{\boldsymbol{\theta}}_n^k\right)\left(\boldsymbol{\theta}_k - \overline{\boldsymbol{\theta}}_n^k\right)' |\mathbf{y}, M_k\right]$ is a consistent estimator of $(\mathbf{H}_n^k)^{-1}$; see Li et al. (2020). $\mathbf{V}_n^k$ can be directly calculated from Markov chain Monte Carlo (MCMC) samples. In (34), $\overline{\boldsymbol{\Omega}}_n^k$ is in fact a robust choice of $\widehat{\mathbf{B}}_n$. Li et al. (2020) used

$$\overline{\boldsymbol{\Omega}}_n^k = \frac{1}{n}\sum_{t=1}^{n}\sum_{\tau=1}^{n} \mathbf{s}_t\left(\overline{\boldsymbol{\theta}}_n^k\right)\mathbf{s}_\tau\left(\overline{\boldsymbol{\theta}}_n^k\right)' k\left(\frac{t-\tau}{\gamma_n}\right), \tag{35}$$

which is a heteroskedasticity and autocorrelation consistent (HAC) estimator of $\mathbf{B}_n^k$, where $k(\cdot)$ is a kernel function and γ_n is the bandwidth; see Newey and West (1987) and Andrews (1991) for more details.

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

$$E_{\mathbf{y}}\left[\text{DIC}_M(k) + 2C\right] = Risk(M_k, d_1) + o(1). \tag{36}$$

where $C = \int \ln \mathbf{g}(\mathbf{y}_{rep})\, \mathbf{g}(\mathbf{y}_{rep})\, \mathrm{d}\mathbf{y}_{rep}$ is a constant that is independent on model. If the candidate model M_k is correctly specified or a good approximation to DGP, $\text{DIC}_M(k)$

becomes

$$\text{DIC}_M(k) = -2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) + 2P_D^k \text{ with } P_D^k = \int 2\left[\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) - \ln p(\mathbf{y}|\boldsymbol{\theta}_k, M_k)\right] \mathrm{d}\boldsymbol{\theta}_k. \tag{37}$$

TIC is defined by

$$\text{TIC}(k) = -2\ln p\left(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n^k, M_k\right) + 2P_T^k \text{ with } P_T^k = -\mathbf{tr}\left\{\bar{\boldsymbol{\Omega}}_n\left(\widehat{\boldsymbol{\theta}}_n^k\right)\widehat{\mathbf{H}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n^k\right)\right\}. \tag{38}$$

Li et al. (2020) established the relationship between $\text{TIC}(k)$ and $\text{DIC}_M(k)$ by showing that

$$E_{\mathbf{y}}\left[\text{DIC}_M(k) + 2C\right] = E_{\mathbf{y}}\left[\text{TIC}(k) + 2C\right] + o(1) = Risk\left(M_k, d_1\right) + o(1). \tag{39}$$

Hence, $\text{DIC}_M(k)$ can be explained as Bayesian version of $\text{TIC}(k)$. When the candidate model is correctly specified or a good approximation to DGP, it was shown in Li et al. (2025) that

$$E_{\mathbf{y}}\left[\text{DIC}_M(k) + 2C\right] = E_{\mathbf{y}}\left[\text{AIC}(k) + 2C\right] + o(1) = Risk\left(M_k, d_1\right) + o(1), \tag{40}$$

where

$$\text{AIC}(k) = -2\ln p(\mathbf{y}|\widehat{\boldsymbol{\theta}}_n^k, M_k) + 2P_k, \tag{41}$$

with P_k being the number of parameters in M_k.

4.2. *Information criterion for comparing misspecified models*

It should be noted that these penalty-based information criteria generally comprise two parts. The first part involves evaluating the log-likelihood at the certain point estimators, which measures the model fit. The second part is the penalty term, which measures the model complexity. What is more, these information criteria are in fact asymptotic unbiased estimators of the corresponding statistical decision risks.

Follow the same logic, we now develop two new information criteria that can be used to estimate $Risk(M_k, d_2)$ and $Risk(M_k, d_3)$. For completeness, we also state the corresponding result of $Risk(M_k, d_1)$ of Li et al. (2020).

When the misspecification is considered, we define three information criteria for model M_k as

$$\text{IC}_1(k) = -2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) + 2P_k^1,$$

$$\text{IC}_2(k) = -2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) + 2P_k^2,$$
$$\text{IC}_3(k) = -2\ln p(\mathbf{y}|\overline{\boldsymbol{\theta}}_n^k, M_k) + 2P_k^3,$$

where

$$P_k^1 = n\mathbf{tr}\left[\overline{\boldsymbol{\Omega}}_n^k \mathbf{V}_n^k\right],$$
$$P_k^2 = (P_k^1 + P_k \ln 2)/2,$$
$$P_k^3 = P_k^1/2 + \ln\left(\left|n\overline{\boldsymbol{\Omega}}_n^k \mathbf{V}_n^k + I_P\right|\right)/2 -$$
$$\mathbf{tr}\left[\left(\overline{\boldsymbol{\Omega}}_n^k - \left(n\mathbf{V}_n^k\right)^{-1}\right)\left(\left(n\mathbf{V}_n^k\right)^{-1} + \left[\left(n\mathbf{V}_n^k\right)\overline{\boldsymbol{\Omega}}_n\left(n\mathbf{V}_n^k\right)\right]^{-1}\right)^{-1}\right]/2,$$

with P_k being the number of parameters in M_k and $\overline{\boldsymbol{\Omega}}_n^k$ being a HAC estimator of $\mathbf{B}_n^k$ defined in (35).

Note that IC_1 is DIC_M in Li et al. (2020), while IC_2 and IC_3 are new to the literature. Li et al. (2020) showed that $\text{IC}_1(k)$ is an asymptotic unbiased estimator of the risk associated with the plug-in predictive distribution when model M_k is potentially misspecified. IC_2 estimates the risk associated with the Bayesian predictive distribution. IC_3 estimates the risk associated with the sandwich Bayesian predictive distribution. In the following, for convenience of description, we use IC_1 to stand for DIC_M.

Consistent with the existing information criteria such as AIC, TIC and DIC, our new information criteria IC_2 and IC_3 consist of two parts: the model fitness and penalty for model complexity. The penalty terms in IC_2 and IC_3 are proposed to capture 'complexity' under the Bayesian predictive distribution and the sandwich Bayesian predictive distribution. In fact, the following theorem guarantees that $\text{IC}_2(k)$ and $\text{IC}_3(k)$ are asymptotic unbiased estimators for $Risk(M_k, d_2)$ and $Risk(M_k, d_3)$.

THEOREM 6: *Under Assumptions 1-9, we have,*

$$E_{g(\mathbf{y})}(IC_2(k)) = Risk\left(M_k, d_2\right) + o(1),$$
$$E_{g(\mathbf{y})}(IC_3(k)) = Risk\left(M_k, d_3\right) + o(1).$$

REMARK 9: When the model is correctly specified, IC_3 reduces to IC_2. That is because IC_2 is based on the Bayesian predictive distribution, while IC_3 is based on the sandwich

Bayesian predictive distribution. When the model is correctly specified, the two predictive distributions are asymptotic equivalent and hence, their corresponding information criteria are the same.

REMARK 10: Based on $Risk(M_k, d_1)$, $Risk(M_k, d_2)$ and $Risk(M_k, d_3)$, we can use IC_1, IC_2 and IC_3 to do model selection. Let the corresponding optimal model be k_1^*, k_2^* and k_3^*. Model k_1^* has the lowest prediction risk under the plug-in predictive distribution. Model k_2^* has the lowest prediction risk under the Bayesian predictive distribution. Model k_3^* has the lowest prediction risk under the sandwich Bayesian predictive distribution. It should be noted that k_1^*, k_2^* and k_3^* may be different in practice.

COROLLARY 7: Let the optimal decision under risk $Risk(M_k, d_2)$ and $Risk(M_k, d_3)$ be k_2^* and k_3^*, respectively. By Corollary 4 and Theorem 6,

$$\lim_{n\to+\infty} Risk(M_{k_2^*}, d_2) \geq \lim_{n\to+\infty} Risk(M_{k_3^*}, d_3).$$

REMARK 11: Therefore, the risk associated with the sandwich predictive posterior cannot be higher than that with the regular Bayesian posterior. This corollary again confirms the importance of the sandwich posterior distribution of Müller (2013).

REMARK 12: It should be noted that our goal is not simply to choose a 'best' model under one predictive distribution. Based on $Risk(M_k, d_1)$, $Risk(M_k, d_2)$ and $Risk(M_k, d_3)$, we can get three 'optimal' models k_1^*, k_2^* and k_3^*, which can be estimated by IC_1, IC_2 and IC_3. By comparing $Risk(M_{k_1^*}, d_1)$, $Risk(M_{k_2^*}, d_2)$, $Risk(M_{k_3^*}, d_3)$, which are estimated by $\text{IC}_1(k_1^*)$, $\text{IC}_2(k_2^*)$ and $\text{IC}_3(k_3^*)$, we can further decide which predictive distribution is the best for the purpose of prediction. We then obtain the optimal model and the corresponding optimal predictive distribution from all $3 \times K$ combinations of K candidate models and the three different predictive distributions. Hence, from a predictive viewpoint, more information can be obtained from our model selection framework. This is the important advantage of our proposed method compared with other existing popular information criteria.

REMARK 13: Table I lists and compare alternative information criteria. We also list the estimation methods and the predictive distribution that these information criteria based on, as well as whether or not they need to assume the candidate model is correctly specification or at least a good approximation to the true data generating process.

TABLE I

ALTERNATIVE INFORMATION CRITERIA

	Estimation Method	Specification	Predictive Distribution	Literature
AIC	MLE	Correct	Plug-in Predictive Distribution	Akaike (1974)
TIC	MLE	Misspecified	Plug-in Predictive Distribution	Takeuchi (1976)
DIC	Posterior Mean	Correct	Plug-in Predictive Distribution	Spiegelhalter et al. (2002)
DIC_L	Posterior Mean	Correct	Plug-in Predictive Distribution	Li et al. (2020)
IC_1/DIC_M	Posterior Mean	Misspecified	Plug-in Predictive Distribution	Li et al. (2020)
IC_2	Posterior Mean	Misspecified	Bayesian Predictive Distribution	New
IC_3	Posterior Mean	Misspecified	Sandwich Predictive Distribution	New

5. SIMULATION STUDIES

We now design two simulation studies to check the performance of the new criteria. In both studies, we compare misspecified models. In the first simulation study, we use the polynomial regression to fit a nonlinear model. In the second simulation study, we try to choose a 'better' model between the logit model and the probit model while the true model is a mixture of logit and probit. We also use other well-known criteria as benchmarks.

5.1. *Polynomial Regression*

In this subsection, we design a simple experiment to compare alternative model selection criteria when the true DGP is not included in the set of candidate models. In other words, all candidate models are misspecified. Following Li et al. (2020), we generate data from the following model

$$y_i = \ln(1 + 46x_i) + e_i,\ e_i \sim N(0,1), i = 1, \ldots, n$$

where $x_i = 0.7(i-1)/n$ which is fixed under repeated sampling by design. In practice, researchers do not know the functional form. Suppose the following set of polynomial regressions is considered,

$$M_k : y_i = \sum_{j=0}^{k-1} \beta_{k,j+1} x_i^j + u_i$$

where $k = 1, \ldots, \left\lfloor n^{1/3} \right\rfloor$ and u_i is assumed to be $N\left(0, \sigma^2\right)$. When $k \to \infty$ as $n \to \infty$, the polynomial regression is related to the sieve estimator that uses progressively more complex models to estimate an unknown function as more data becomes available. In our experiment, we estimate and compare all the candidate models $\left\{M_k, k = 1, \ldots, \left\lfloor n^{1/3} \right\rfloor\right\}$. In M_k, $\sum_{j=0}^{k-1} \beta_{k,j+1} x_i^j$ is used to approximate $\ln(1 + 46x_i)$. Let $\boldsymbol{\beta}_k = (\beta_1, \ldots, \beta_k)'$ so that $\boldsymbol{\theta}_k = \left(\boldsymbol{\beta}_k', \sigma^2\right)$ and the number of parameters is $k + 1$. Let $\mathbf{x}^j = \left(x_1^j, x_2^j, \ldots, x_n^j\right)'$, $\mathbf{X}_k = \left(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{k-1}\right)$, and $\mathbf{X} = \left(\mathbf{x}^0, \mathbf{x}^1, \ldots, \mathbf{x}^{\left\lfloor n^{1/3} \right\rfloor - 1}\right)$.

Three different sample sizes are considered, $n = 100, 500, 1000$. For each candidate model M_k, we obtain the MLE of $\boldsymbol{\theta}_k$, denoted by $\widehat{\boldsymbol{\theta}}_k = \left(\hat{\boldsymbol{\beta}}_k', \hat{\sigma}^2\right)$, and then calculate AIC and TIC. $\widehat{\boldsymbol{\theta}}_k$, which is also the OLS estimate, has a closed-form expression for this model. The following g-prior is used for $\boldsymbol{\theta}_k$ do conduct the Bayesian analysis,

$$\pi\left(\sigma^2\right) \propto \frac{1}{\sigma^2}, \quad \boldsymbol{\beta}_k \sim N\left(\boldsymbol{\beta}_{k,0}, g\sigma^2 \left(\mathbf{X}_k' \mathbf{X}_k\right)^{-1}\right),$$

where $g = n$ denotes the unit information prior of Kass and Wasserman (1995) in the normal regression case. The posterior mean and the posterior variance of θ_k are

$$\begin{aligned}
E\left(\boldsymbol{\beta}_k \mid \mathbf{y}, \mathbf{X}\right) &= \frac{g}{g+1}\left(\frac{\boldsymbol{\beta}_{k,0}}{g} + \hat{\boldsymbol{\beta}}_k\right), \\
E\left(\sigma^2 \mid \mathbf{y}, \mathbf{X}\right) &= \frac{s^2 + \frac{1}{g+1}\left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k,0}\right)' \mathbf{X}_k' \mathbf{X}_k \left(\hat{\boldsymbol{\beta}}_k - \boldsymbol{\beta}_{k,0}\right)}{n-2}, \\
\operatorname{Var}\left(\boldsymbol{\beta}_k \mid \mathbf{y}, \mathbf{X}\right) &= \frac{g}{g+1}\left(\mathbf{X}_k' \mathbf{X}_k\right)^{-1} E\left(\sigma^2 \mid \mathbf{y}, \mathbf{X}\right), \\
\operatorname{Var}\left(\sigma^2 \mid \mathbf{y}, \mathbf{X}\right) &= \frac{2E\left(\sigma^2 \mid \mathbf{y}, \mathbf{X}\right)^2}{n-4}, \\
\operatorname{Cov}\left(\boldsymbol{\beta}_k, \sigma^2 \mid \mathbf{y}, \mathbf{X}\right) &= 0.
\end{aligned}$$

These closed-form expressions are used to calculate DIC, DIC_L, IC_1, IC_2 and IC_3.

We replicate the simulation experiment for 1000 times. In every experiment, we simulate $\mathbf{y}$ from the true model and calculate seven criteria for each candidate model M_k with $k = 1, \ldots, \left\lfloor n^{1/3} \right\rfloor$. Each of the seven criteria is used to select a best model (call it M_{k^*} may differ across different criteria), we record this model and the corresponding IC.

Note that for AIC and TIC, we take MLE under their best model as the estimator and use the plug-in density $p(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_{k^*}, M_{k^*})$ to predict new data. For DIC, DIC_M and IC_1, we use the plug-in predictive distribution $p(\mathbf{y}_{rep}|\bar{\boldsymbol{\theta}}_{k^*}, M_{k^*})$ under the best model M_{k^*} to predict new data. For IC_2, we use the regular Bayesian predictive distribution $p(\mathbf{y}_{rep}|\mathbf{y}, M_{k^*})$ under the best model M_{k^*} to predict new data. For IC_3, we use the sandwich predictive distribution $p^s(\mathbf{y}_{rep}|\mathbf{y}, M_{k^*})$ under the best model M_{k^*} to predict new data. Then we replicate the experiment 1000 time and estimate each of the seven risk functions using the average of the corresponding ICs.

Table II reports the relative frequencies of the selected models by each of seven criteria (namely AIC, TIC, DIC, DIC_L, IC_1, IC_2, IC_3), the average values of k^*, and the average value of the estimated risks for each of seven criteria, all across 1000 replications.[12]

Several interesting results can be found in Table II. First, the models selected by the BIC tend to be more parsimonious than those selected by other criteria. This result is not surprising as BIC has a larger penalty term than other criteria. Second, the average k^*s selected by AIC, TIC, DIC, DIC_L and IC_1 are very similar, suggesting that they tend to select the same model. This is not surprising because AIC, TIC, DIC, DIC_L and IC_1 all use the plug-in predictive distribution to calculate the predictive loss. Third, IC_2 tends to choose more complex model than all the criteria based on the plug-in predictive distribution while IC_3 tends to choose even more complex model then IC_2. Of course the complex model is closer to the true model. Fourth, as the sample size increases, the average k^*s selected by all criteria tend to increase.

Now let us focus on the estimated risk of IC_1, IC_2 and IC_3. IC_3 has a smaller risk than IC_2, and IC_2 has a smaller risk than IC_1. Results obtained from this Monte Carlo study indicate that if one's objective is to get a best prediction for future data, we should not only consider how to choose the 'best' model and estimator, but also consider what predictive distribution we should use.

[12] We report $(\widehat{Risk}/n - 1 - \ln(2\pi)) \times 10^3$ instead of $\widehat{Risk}$ to better highlight differences in the estimated risk functions under different criteria.

TABLE II

SIMULATION RESULTS FOR THE FIRST EXPERIMENT

	AIC	TIC	DIC	DIC_L	IC_1/DIC_M	IC_2	IC_3
Relative frequency of the polynomial order selected by alternative ($n = 100$)							
$k = 2$	0.064	0.061	0.065	0.063	0.052	0.042	0.041
$k = 3$	0.510	0.491	0.504	0.510	0.468	0.451	0.441
$k = 4$	0.426	0.448	0.431	0.427	0.480	0.507	0.518
Relative frequencies of the polynomial order selected by different criteria ($n = 500$)							
$k = 3$	0.062	0.062	0.063	0.062	0.059	0.044	0.042
$k = 4$	0.348	0.334	0.341	0.346	0.327	0.287	0.283
$k = 5$	0.318	0.321	0.323	0.318	0.322	0.323	0.321
$k = 6$	0.167	0.174	0.168	0.168	0.180	0.200	0.200
$k = 7$	0.105	0.109	0.105	0.106	0.112	0.146	0.154
Average value of the estimated risk under alternative IC ($n = 100$)							
Risk	49.436	48.602	50.573	50.935	44.418	35.901	33.917
Standard Error	(4.732)	(4.746)	(4.708)	(4.711)	(4.743)	(4.722)	(4.724)
Average value of the estimated risk under alternative IC ($n = 500$)							
Risk	12.128	12.020	12.131	12.178	11.704	8.330	8.167
Standard Error	(1.989)	(1.990)	(1.988)	(1.989)	(1.990)	(1.990)	(1.990)

5.2. *The mixture of logit and probit*

The logit model and the probit model are widely used for discrete choices. In the second experiment, we simulate data from the mixture of the two models and use alternative IC to choose between the logit model and the probit model.

Suppose $\mathbf{y} = (y_1, y_2, ..., y_n)'$ be a vector of dependent variables, y_i takes values 0 or 1 for $i = 1, 2, ..., n$, the independent variable matrix $X = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N]'$ where $\mathbf{x}_i$ is a $P \times 1$ vector. The probability of $y_i = 1$ conditional on $\mathbf{x}_i$ is

$$P(y_i = 1|\mathbf{x}_i, \beta) = F(\mathbf{x}_i'\beta), \tag{42}$$

where β is a $P \times 1$ vector and $(y_i, \mathbf{x}_i)$ are identical and independent distributed. If we choose $F(\mathbf{x}_i'\beta) = \Phi(\mathbf{x}_i'\beta)$ with $\Phi(\cdot)$ be the CDF of standard normal distribution, (42) is the probit model. If choosing $F(\mathbf{x}_i'\beta)$ be the CDF of logistic distribution

$$F(\mathbf{x}_i'\beta) = \frac{\exp(\mathbf{x}_i'\beta)}{1 + \exp(\mathbf{x}_i'\beta)},$$

(42) becomes the logit model. The latent variable representation of (42) is as follows

$$y_i = \mathbf{I}\left(z_i > 0\right), z_i = \mathbf{x}_i'\beta + \varepsilon_i,$$

where $\mathbf{I}(\cdot)$ is the indicator function, the density function of ε_i is $f(\varepsilon_i) = \phi(\varepsilon_i)$ with $\phi(\cdot)$ be the PDF of the standard normal distribution for the probit model, and

$$f(\varepsilon_i) = \frac{\exp(\varepsilon_i)}{(1 + \exp(\varepsilon_i))^2}$$

for the logit model. For model comparison, we denote the probit and logit model as Model 1 and Model 2, named by M_1 and M_2, respectively.

We simulate from a mixture of the probit model and the logit model, so that both M_1 and M_2 are misspecified. We generate i.i.d. data from the following model

$$\begin{aligned}
&\varepsilon_{1i} \sim N(0,1),\ \varepsilon_{2i} \sim logistic(0,1),\ U \sim U(0,1),\\
&\varepsilon_i = \mathbf{I}(U \le q)\varepsilon_{1i} + \mathbf{I}(U > q)\varepsilon_{2i},\\
&y_i = \mathbf{I}(z_i > 0), z_i = \mathbf{x}_i'\beta + \varepsilon_i,
\end{aligned}$$

where $q \in [0,1]$ is a given parameter. For simplicity we write $\varepsilon_i \sim q \times N(0,1) + (1-q) \times logistic(0,1)$.

In this model, to simulate ε_i, we generate a random number ε_{1i} from $N(0,1)$ and a random number ε_{2i} from the standard logistic distribution. Then we let $\varepsilon_i = \varepsilon_{1i}$ with probability q and let $\varepsilon_i = \varepsilon_{2i}$ with probability $1-q$. If we specify $q = 1$, i.e., $\varepsilon_i \sim N(0,1)$, then we get a probit model (we denote it as M_1). If we specify $q = 0$, i.e., $\varepsilon_i \sim logistic(0,1)$, then we get a logit model (we denote it as M_2). Thus, we simulate data from the mixture of probit and logit where parameter q controls the proportions of probit and logit. When q is closed to 1, the model is closer to a probit model than to a logit model.

Now suppose we do not know the true DGP. We choose between M_1 and M_2. Thus, we use the seven criteria to make model selection. To compare M_1 and M_2, we need to estimate them first. The calculation of AIC and TIC requires QMLE, which is easily obtained by a standard statistical software. However, the other information criteria need the posterior mean, which is harder to get.

Albert and Chib (1993) proposed a Gibbs Sampling algorithm for the probit model based on data augmentation of Tanner and Wong (1987). It draws samples from the joint posterior distribution of the parameters and the latent variables. The latent variables can be drawn form a conditionally normal distribution since they follow a linear model of parameters with a normal error term. Zens et al. (2023b) applied marginal data augmentation (Liu and Wu (1999)) to boost the convergence of the Gibbs Sampling algorithm for the probit model. For the logit model, the latent variable follows a linear model with a logistic error term. Holmes and Held (2006) used the scale mixture normal representation of the logistic error with the Kolmogorov-Smirnov distribution. Polson et al. (2013) proposed a new mixture representation of the logistic error with Pólya-Gamma distribution that can largely improve the efficiency of the Gibbs Sampling algorithm. Zens et al. (2023b) proposed a ultimate Pólya-Gamma (UPG) samplers with marginal data augmentation to further improve the efficiency of the Gibbs Sampling algorithm for the logit model. In this paper, we use UPG for the probit model and the logit model, which is implemented by the UPG package in R. To conduct the Bayesian analysis, we specify a vague prior distribution

$$\beta \sim N(0, 10 \times I_k),$$

We use the UPG package in R and draw 11000 MCMC samples from the posterior distribution. The first 1000 is used as the burn-in sample, and the next 10,000 iterations is collected as effective MCMC draws. With the posterior samples, we can obtain the posterior mean $\bar{\beta}$ and DIC, DIC_L, IC_1, IC_2 and IC_3.

We simulate for $q = 0, 0.1, 0.2, ..., 0.9, 1$. For each q, we simulate data with sample size $n = 500$ and calculate AIC, TIC, DIC, DIC_L, IC_1, IC_2, IC_3. Then replicate this experiment for 1000 times. The performance of these criteria is compared based on these 1000 replications. We calculate the risks of every criterion using the same method as in Section 5.1. Table III reports the risks of all seven criteria, the corresponding standard errors are

reported in the parentheses. AIC, TIC, DIC, DIC_L and IC_1 have similar risks, while the risk of IC_3 is lower than those of IC_1 and IC_2 for every q. This indicates that the sandwich predictive distribution can reduce the risk of statistical decision. Compared with other criteria, even if IC_3 chooses the same model, we can use the sandwich predictive distribution to improve the prediction.

TABLE III

THE AVERAGE RISKS UNDER DIFFERENT CRITERIA

Criteria	AIC	TIC	DIC	DIC_L	$\mathrm{IC}_1/\mathrm{DIC}_M$	IC_2	IC_3
$q=0$	520.071	520.066	520.066	520.104	520.106	519.172	519.159
s.e.	(0.651)	(0.651)	(0.651)	(0.651)	(0.651)	(0.651)	(0.651)
$q=0.1$	508.725	508.719	508.720	508.760	508.760	507.826	507.813
s.e.	(0.663)	(0.664)	(0.663)	(0.663)	(0.664)	(0.663)	(0.664)
$q=0.2$	500.129	500.137	500.121	500.163	500.177	499.237	499.223
s.e.	(0.654)	(0.654)	(0.654)	(0.654)	(0.654)	(0.654)	(0.654)
$q=0.3$	489.133	489.162	489.124	489.167	489.202	488.251	488.236
s.e.	(0.680)	(0.680)	(0.680)	(0.680)	(0.680)	(0.680)	(0.680)
$q=0.4$	477.076	477.103	477.064	477.109	477.142	476.193	476.177
s.e.	(0.667)	(0.667)	(0.667)	(0.667)	(0.667)	(0.667)	(0.667)
$q=0.5$	463.546	463.568	463.537	463.583	463.611	462.663	462.645
s.e.	(0.690)	(0.690)	(0.690)	(0.690)	(0.690)	(0.690)	(0.690)
$q=0.6$	453.332	453.367	453.319	453.366	453.406	452.453	452.435
s.e.	(0.693)	(0.693)	(0.693)	(0.693)	(0.693)	(0.693)	(0.693)
$q=0.7$	437.997	438.023	437.977	438.023	438.054	437.110	437.092
s.e.	(0.716)	(0.716)	(0.716)	(0.716)	(0.716)	(0.716)	(0.716)
$q=0.8$	424.242	424.225	424.224	424.268	424.258	423.335	423.314
s.e.	(0.700)	(0.700)	(0.700)	(0.700)	(0.700)	(0.700)	(0.700)
$q=0.9$	408.538	408.481	408.517	408.559	408.511	407.609	407.589
s.e.	(0.686)	(0.687)	(0.687)	(0.687)	(0.687)	(0.687)	(0.687)
$q=1$	393.119	392.961	393.096	393.136	392.990	392.139	392.118
s.e.	(0.653)	(0.654)	(0.654)	(0.653)	(0.654)	(0.653)	(0.654)

6. EMPIRICAL STUDIES

6.1. *Discrete choice models*

In the empirical research, discrete choice models have been widely used. In this section, we consider a model comparison between a binary probit model (M_1) and a binary logit model (M_2). The data set is the female labor force participation from the US Panel Study of Income, including a binary dependent variable takes the value of 1 if the woman is participating in the labor force, the number of children under the age of 5, the number of children between 6 and 18 years, a standardized age index, two binary indicators capturing whether a college degree was obtained by the wife and the husband, the expected log wage of the woman, the logarithm of family income exclusive of the income of the woman. There are 753 observations in the data set. For more details about the data, see Zens et al. (2023a). Then there are 8 parameters in both models including the intercepts.

To obtain MCMC output, we first specify a vague prior distribution for parameters as

$$\beta \sim N(0_{k\times 1}, \lambda \times I_k),$$

where $\lambda = 100$ in both models. Then we use more informative priors with $\lambda = 10$ or 1. To draw MCMC samples, we use the same method as that in Section 5.2. We draw 510,000 random draws from the joint posterior distributions of parameters in each model. The first 10,000 is used as the burn-in sample, and the next 500,000 iterations is collected as effective observations. Hence, there are 500,000 effective draws.

To compare the two models, based on 500,000 effective draws, we calculate AIC, TIC, DIC, DIC_L, IC_1, IC_2 and IC_3 for two candidate models under different priors.

Table IV reports the model selection results under various prior. Several interesting results may be found in the table. First, it is unsurprising to see AIC and DIC take similar values in all cases as they are asymptotically equivalent as shown in Li et al. (2025). However, we are surprised to see AIC, DIC, DIC_L, IC_3 take similar values in all cases because while AIC and DIC assume the models are correctly specified while DIC_M and IC_3 allow model misspecification.

Second, TIC takes very different values from AIC in all cases, suggesting both models are misspecified, and hence TIC is more applicable. Interestingly, AIC suggests M_2 is preferred, TIC suggests M_1 is preferred.

TABLE IV

MODEL SELECTION RESULTS FOR MODEL 1 AND 2 UNDER DIFFERENT PRIOR

λ	AIC	TIC	DIC	DIC_L	IC_1/DIC_M	IC_2	IC_3
				M_1			
1	921.3899	948.9662	921.2682	921.2658	948.8300	932.6584	921.6041
10	921.3899	948.9662	921.3975	921.4283	949.0664	932.7751	921.6842
100	921.3899	948.9662	921.3958	921.4302	948.8273	932.6563	921.6876
				M_2			
1	921.2659	950.7182	920.9610	920.9141	950.7100	933.5714	921.5698
10	921.2659	950.7182	921.2944	921.4513	950.9292	933.6507	921.7726
100	921.2659	950.7182	921.3576	921.5378	951.0234	933.7023	921.8164

Third and most importantly, IC_3 takes much lower values than IC_2 that in turn takes much lower values than IC_3. This is consistent with our theoretical results. Since IC_3 is smaller than IC_1 and IC_2, it suggests the sandwich predictive distribution leads to the smallest KL losses in all case. According to IC_3, M_1 is preferred to M_2 when a moderately vague or a vague prior is used (i.e., $\lambda = 10, 100$). However, M_2 is preferred to M_1 when an informative prior is used (i.e., $\lambda = 1$),

6.2. *SV models*

Stochastic volatility (SV) models have been found very useful for pricing derivative securities and modeling time-varying volatility. The discrete-time basic log-normal SV model is composed of two equations. One is the measurement equation, the other is state equation where the logarithmic volatility is the state variable. The state equation is assumed to follow an AR(1) model. The basic log-normal SV model is of the form:

$$y_t = \exp(h_t/2)u_t,\ u_t \sim N(0,1), t = 1, ..., n,$$

$$h_t = \mu + \phi(h_{t-1} - \mu) + \tau v_t, v_t \sim N(0,1), h_0 = \mu,$$

where y_t is the continuously compounded return, h_t is the unobserved log-volatility, u_t and v_t are serially independent for all t, $corr(u_t v_s) = 0$ for any t, s. In this paper, we denote this model M_1.

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

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

An important and well documented empirical feature in many financial time series is the leverage effect. Following Yu (2005), the leverage effect SV model allows for correlation between the two error terms, that is, $corr(u_t, v_s) = \rho$. In this model, ρ captures the leverage effect if $\rho < 0$. We denote this model M_2 and specify the prior distribution of ρ as $\rho \sim Uniform(-1, 1)$.

SV models are difficult to estimate by ML, and hence, it is hard to calculate AIC and TIC. Our goal is to compare the two models using DIC_L, IC_1, IC_2 and IC_3. Note that both models are nonlinear non-Gaussian state-space models, the state variable h_t is latent. Thus, the likelihood function $p(\mathbf{y}|\theta)$ is not available in close-form . That is why a popular estimation and inferential method is Bayesian MCMC.

The dataset consists of 945 daily mean-corrected returns on Pound/Dollar exchange rates, covering the period between 01/10/81 and 28/06/85. For MCMC, after a burn-in period of 10,000 iterations, we save every 20th value for the next 100,000 iterations to get 5,000 effective draws. The same dataset was used in Kim et al. (1998) and Meyer and Yu (2000).

Table V gives the posterior mean and the posterior standard error of parameters in the basic SV model (M_1) and the leverage SV model (M_2). Also note that in M_1 and M_2, the posterior mean and the posterior standard error of μ, ϕ and τ are all similar. Moreover, the posterior mean of ρ is very close to zero, relative to its posterior standard error. This indicates that the leverage effect may be no significant. From the point of simplification, the basic SV model may be a better choice.

Table VI reports $2P_L$, $2P_1$, $2P_2$, $2P_3$, DIC_L, IC_1, IC_2, and IC_3. First, all four criteria choose the basic SV model (M_1), which coincides with our analysis in Table V. Judged by the difference among DIC_L, IC_1, IC_2 and IC_3 and the difference among $2P_L$, $2P_1$, $2P_2$,

TABLE V

POSTERIOR MEAN AND STANDARD ERROR OF PARAMETERS IN M_1 AND M_2

	M_1		M_2	
Parameter	Mean	SE	Mean	SE
μ	-0.7158	0.3008	-0.6711	0.3529
ϕ	0.9767	0.0145	0.9771	0.0143
ρ	NA	NA	-0.0104	0.1381
τ	0.1771	0.0161	0.1806	0.0144

$2P_3$, we can tell that the difference of information criteria is mainly caused by the penalty term. Because the leverage SV model M_2 has one more parameter ρ, its penalty terms are larger than those of M_1. The extra parameter ρ does not improve the model fitting a lot. That explains why a parsimonious model is selected.

Moreover, among IC_1, IC_2 and IC_3, IC_3 of M_1 is the smallest. This observation suggests that we prefer not only the basic SV model (M_1) but also the sandwich Bayesian predictive distribution for the purpose of predicting future data.

TABLE VI

MODEL SELECTION RESULTS FOR M_1 AND M_2

Model	$2P_L$	DIC_L	$2P_1$	IC_1/DIC_M	$2P_2$	IC_2	$2P_3$	IC_3
M_1	4.935	1843.784	8.658	1847.507	6.409	1845.257	6.088	1844.936
M_2	6.510	1845.406	10.451	1849.346	7.998	1846.894	7.710	1846.606

7. CONCLUSION

It is well known that in Bayesian literature, when the model is misspecified, the posterior distribution still has an asymptotic normal distribution which centered at the maximum likelihood estimator (MLE) with Hessian information matrix which is in general, different than the "sandwich" covariance matrix. In a recent literature, Müller (2013) showed that due to this discrepancy between the Hessian information matrix and sandwich covariance matrix, an artificial normal posterior centered at MLE with sandwich covariance matrix

(sandwich posterior, hereafter) can yield lower asymptotic frequentist risk than the original normal posterior. On the basis of these two different posteriors, from predictive viewpoint, three are three different predictive distributions for candidate use, that is, Plug-in predictive distribution, Bayesian predictive distribution, and Müller's Bayesian predictive distribution based on the sandwich posterior distribution.

In this paper, the main contributions are least threefold. First, from predictive viewpoint, we investigate the theoretical properties how these three predictive distributions work and which can actually outperform best in a variety of settings. On the basis of Kullback-Leibler (KL) loss function, we show that the sandwich Bayesian predictive distribution also can yield lower asymptotic risk than the standard posterior distribution. Furthermore, we give the conditions from the asymptotic risk that the sandwich Bayesian predictive distribution is better or not than the plug-in predictive distribution. Second, based on the Bayesian predictive distribution and sandwich Bayesian predictive distribution, we proposed two important information criterion for comparing misspecified models which can be unbiased estimators for the risks based on corresponding predictive distributions. Third, we established the relationship between the propose information criterion and the existing information criterion such as the popular AIC, TIC, and DIC, etc. At last, we illustrate the proposed new information criteria using some real studies in economics and finance.

APPENDIX

A.1. *Notations*

$:=$	definitional equality	$\overleftrightarrow{\boldsymbol{\theta}}_n$	posterior mode
$o(1)$	tend to zero	$\widehat{\boldsymbol{\theta}}_n$	QML estimator
$o_p(1)$	tend to zero in probability	$\boldsymbol{\theta}_n^p$	pseudo true parameter
$\xrightarrow{p}$	converge in probability	$\overline{\boldsymbol{\theta}}_n$	posterior mean

A.2. *Proof of Theorem 1*

We provide a proof sketch in this Appendix, details are given in the Supplement. Denote

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

These three lemma are useful to prove our result.

LEMMA 8: *Under Assumptions 1-9,* $\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p \xrightarrow{p} 0$, $\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p \xrightarrow{p} 0$.

LEMMA 9: *Under Assumptions 1-9, the following asymptotic expansions hold:*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) = -\mathbf{H}_n^{-1}\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}} + RT_n^0(\mathbf{y}),$$

$$\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) = (-2\mathbf{H}_n)^{-1}\left(\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}} + \frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}}\right) + RT_n^1(\mathbf{y}, \mathbf{y}_{rep}),$$

$$\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right) = -\mathbf{H}_n^{-1}\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}} + RT_n^2(\mathbf{y}),$$

where $E|RT_n^0(\mathbf{y})|^2 = o(1)$, $E|RT_n^1(\mathbf{y}, \mathbf{y}_{rep})|^2 = o(1)$, $E|RT_n^2(\mathbf{y})|^2 = o(1)$.

LEMMA 10: *Under Assumptions 1-9, the following moment conditions hold*

$$E\left\|\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right\|^4 \leq C,\ E\left\|\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right\|^4 \leq C,\ E\left\|\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\right\|^4 \leq C,$$

$$E\left\|\sqrt{n}\bar{\mathbf{s}}(\boldsymbol{\theta}_n^p)\right\|^2 \leq C,\ E\left\|\sqrt{n}(\overline{\mathbf{H}}_n(\boldsymbol{\theta}_n^p) - \mathbf{H}_n)\right\|^2 \leq C.$$

LEMMA 11: *Under Assumptions 1-9,*

$$E\left[\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)'\right] = 2^{-1}\boldsymbol{\Sigma}_n + o(1),$$

where $\boldsymbol{\Sigma}_n = \mathbf{H}_n^{-1}\mathbf{B}_n\mathbf{H}_n^{-1}$.

We are now in the position to prove Theorem 1. By the Laplace approximation (Tierney et al., 1989, Kass et al., 1990) and Lemma 9, we have

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

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

where

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

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

Then we have

$$\begin{aligned}&\ln p(\mathbf{y}_{rep}|\mathbf{y})\\ =&\underbrace{-\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right|\right)}_{T_1} + \underbrace{\left[-nh_N\left(\widetilde{\boldsymbol{\theta}}_n\right) + nh_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]}_{T_2} + RT_n^3.\end{aligned}$$

where $E|RT_n^3| = 0$.

The first term can be approximated by

$$\begin{aligned}T_1 &= -\frac{1}{2}\left(\ln\left|\nabla^2 h_N\left(\widetilde{\boldsymbol{\theta}}_n\right)\right| - \ln\left|\nabla^2 h_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right|\right)\\ &= -\frac{1}{2}\ln|-\mathbf{H}_n - \mathbf{H}_n| + \frac{1}{2}\ln|-\mathbf{H}_n| + RT_n^4 = -\frac{1}{2}P\ln 2 + RT_n^4,\end{aligned} \tag{43}$$

where $E|RT_n^4| = 0$. The second term can be approximated by

$$T_2 = -n\hat{h}_N\left(\widetilde{\boldsymbol{\theta}}_n\right) + n\hat{h}_D\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right) = \ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right) + E_1 + E_2 + RT_n^5, \tag{44}$$

where $RT_n^5 = \ln p\left(\widetilde{\boldsymbol{\theta}}_n\right) - \ln p\left(\overleftrightarrow{\boldsymbol{\theta}}_n\right)$ satisfies $E|RT_n^5| = o(1)$, and

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

We can further decompose E_1 as $E_1 = E_{11} + E_{12}$, where

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

For E_{11}, we have

$$E_{11} = \underbrace{\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)}_{E_{111}} + \underbrace{\frac{1}{2}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)'\frac{1}{n}\frac{\partial^2 \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^*\right)}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_n^p\right)}_{E_{112}}. \tag{45}$$

where $\boldsymbol{\theta}_n^*$ lies between $\widetilde{\boldsymbol{\theta}}_n$ and $\boldsymbol{\theta}_n^p$. By Lemma 11, we can show

$$\begin{aligned}E(E_{111}) &= E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\mathbf{tr}\left[(-2\mathbf{H}_n)^{-1}\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}}\frac{1}{\sqrt{n}}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right)}{\partial \boldsymbol{\theta}'}\right]\right] + o(1)\\ &= \mathbf{tr}\left[(-2\mathbf{H}_n)^{-1}\mathbf{B}_n\right] + o(1) = \frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right] + o(1).\end{aligned} \tag{46}$$

Moreover,

$$E_{112}=\frac{1}{2}\text{tr}\left(\mathbf{H}_n\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)\sqrt{n}\left(\widetilde{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)'\right)+RT_n^6. \tag{47}$$

Then, using Lemma 11, we have

$$E(E_{112})=\frac{1}{2}\text{tr}(\mathbf{H}_n 2^{-1}\boldsymbol{\Sigma}_n)+o(1)=-\frac{1}{4}\mathbf{tr}(\mathbf{B}_n(-\mathbf{H}_n)^{-1})+o(1).$$

Then we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{11})=E(E_{111})+E(E_{112})=\frac{1}{4}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right]+o(1).$$

For E_{12}, we have

$$E_{12}=-\frac{1}{\sqrt{n}}\frac{\partial\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p)}{\partial\boldsymbol{\theta}'}\sqrt{n}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)-\frac{1}{2}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right)'\frac{\partial^2\ln p(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^*)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\left(\overleftrightarrow{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_n^p\right). \tag{48}$$

Taking expectation with respect to both $\mathbf{y}$ and $\mathbf{y}_{rep}$, the first term is exactly 0 because of the independence between $\mathbf{y}$ and $\mathbf{y}_{rep}$. The second term can be treated similarly as $E(E_{112})$. Then we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{12})=\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right]+o(1).$$

Then

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_1)=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{11}+E_{12})=\frac{3}{4}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right]+o(1). \tag{49}$$

By applying the similar method to $E_2=-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)+\ln p\left(\mathbf{y}|\widetilde{\boldsymbol{\theta}}_n\right)$, we get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_2)=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{21}+E_{22})=-\frac{1}{4}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right]+o(1). \tag{50}$$

Recall (44) and we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(T_2)=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]+\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n(-\mathbf{H}_n)^{-1}\right]+o(1).$$

Break $E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]$ into three terms, we get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}\left[\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]=E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]+E_{\mathbf{y}}(E_{31})+E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{32}),$$

where

$$E_{31}=\ln p\left(\mathbf{y}|\boldsymbol{\theta}_n^p\right)-\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right),E_{32}=\ln p\left(\mathbf{y}_{rep}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)-\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}_n^p\right).$$

Following the similar argument of E_{111} and E_{112}, we have

$$E_{\mathbf{y}}(E_{31})=-\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]+o(1),$$
$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(E_{32})=-\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]+o(1).$$

So we have

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

Then we get

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(T_2)=E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]-\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]+o(1). \tag{51}$$

Combining (43) and (51), we have

$$E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)]=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(T_1)+E_{\mathbf{y}}E_{\mathbf{y}_{rep}}(T_2)+o(1)$$
$$=E_{\mathbf{y}}\left[\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]-\frac{1}{2}\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]-\frac{1}{2}P\ln 2+o(1).$$

We finally get the desired result:

$$Risk(d_2)=E_{\mathbf{y}}E_{\mathbf{y}_{rep}}[-2\ln p\left(\mathbf{y}_{rep}|\mathbf{y}\right)]$$
$$=E_{\mathbf{y}}\left[-2\ln p\left(\mathbf{y}|\overleftrightarrow{\boldsymbol{\theta}}_n\right)\right]+\mathbf{tr}\left[\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1}\right]+P\ln 2+o(1).$$

Note that $\overleftrightarrow{\boldsymbol{\theta}}_n=\overline{\boldsymbol{\theta}}_n+O_p(1)$ (see Li et al. (2025)), the first term can be replaced by $E_{\mathbf{y}}\left[-2\ln p\left(\mathbf{y}|\overline{\boldsymbol{\theta}}_n\right)\right]$ without changing the result.

A.3. *Proof of Theorem 2*

Denote

$$\widetilde{\boldsymbol{\theta}}_n^s:=\arg\max_{\boldsymbol{\theta}}\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right)-\frac{n}{2}\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right)'\widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\widehat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}\right),$$

where $\widehat{\boldsymbol{\Sigma}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)$ is a consistent estimator of $\boldsymbol{\Sigma}_n$. By the Laplace approximation,

$$
\begin{aligned}
p^s\left(\mathbf{y}_{rep}|\mathbf{y}\right) &= \int p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right) p^s\left(\boldsymbol{\theta}|\mathbf{y}\right) d\boldsymbol{\theta} \\
&= \left(\frac{1}{2\pi}\right)^{\frac{P}{2}} \left|\frac{\widehat{\boldsymbol{\Sigma}}_n(\widehat{\boldsymbol{\theta}})}{n}\right|^{-\frac{1}{2}} \int \exp\left[\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right) - \frac{n}{2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)\right] d\boldsymbol{\theta} \\
&= \left|\frac{1}{n}\widehat{\boldsymbol{\Sigma}}_n\left(\widehat{\boldsymbol{\theta}}\right)\right|^{-\frac{1}{2}} \left|n\nabla^2 h_N^s\left(\widetilde{\boldsymbol{\theta}}_n^s\right)\right|^{-1/2} \exp\left(-n h_N^s\left(\widetilde{\boldsymbol{\theta}}_n^s\right)\right)\left(1 + O_p\left(\frac{1}{n}\right)\right),
\end{aligned}
$$

where

$$
\begin{aligned}
h_N^s\left(\boldsymbol{\theta}\right) &= -\frac{1}{n}\left(\ln p\left(\mathbf{y}_{rep}|\boldsymbol{\theta}\right) - \frac{n}{2}\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)' \widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\widehat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}\right)\right), \\
\nabla^2 h_N^s\left(\widetilde{\boldsymbol{\theta}}_n^s\right) &= -\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right).
\end{aligned}
$$

So we get the following expansion

$$
\begin{aligned}
\ln p^s\left(\mathbf{y}_{rep}|\mathbf{y}\right) &= -\frac{1}{2}\ln\left|\widehat{\boldsymbol{\Sigma}}_n \nabla^2 h_N^s\left(\widetilde{\boldsymbol{\theta}}_n^s\right)\right| - n h_N^s\left(\widetilde{\boldsymbol{\theta}}_n^s\right) + RT_n^8 \qquad (52) \\
&= L_1 + L_2 + L_3 + \ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right) + RT_n^8, \qquad (53)
\end{aligned}
$$

where

$$
\begin{aligned}
L_1 &= -\frac{1}{2}\ln\left|\widehat{\boldsymbol{\Sigma}}_n\left(\widehat{\boldsymbol{\theta}}_n\right)\left(-\frac{1}{n}\frac{\partial \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n\right)}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'} + \widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\right)\right|, \qquad (54) \\
L_2 &= \ln p\left(\mathbf{y}_{rep}|\widetilde{\boldsymbol{\theta}}_n^s\right) - \ln p\left(\mathbf{y}_{rep}|\widehat{\boldsymbol{\theta}}_n\left(\mathbf{y}_{rep}\right)\right), \qquad (55) \\
L_3 &= -\frac{n}{2}\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n^s\right)' \widehat{\boldsymbol{\Sigma}}_n^{-1}\left(\widehat{\boldsymbol{\theta}}_n\right)\left(\widehat{\boldsymbol{\theta}}_n - \widetilde{\boldsymbol{\theta}}_n^s\right). \qquad (56)
\end{aligned}
$$

By the same argument as in the proof of Theorem 1, we can show that

$$
\begin{aligned}
E(L_1) &= -\frac{1}{2}\ln\left|\mathbf{B}_n\left(-\mathbf{H}_n\right)^{-1} + I_P\right| + o\left(1\right), \qquad (57) \\
E(L_2) &= \mathbf{tr}\left[(-\mathbf{H}_n + \boldsymbol{\Sigma}_n^{-1})^{-1}\mathbf{B}_n\right] + \frac{1}{2}\mathbf{tr}\left[\mathbf{H}_n\mathbf{D}_n\right] - \frac{1}{2}\mathrm{tr}[\mathbf{B}_n(-\mathbf{H}_n^{-1})] + o(1), \qquad (58) \\
E(L_3) &= \mathbf{tr}\left[-\boldsymbol{\Sigma}_n^{-1}(-\mathbf{H}_n + \boldsymbol{\Sigma}_n^{-1})^{-1}\mathbf{B}_n(-\mathbf{H}_n + \boldsymbol{\Sigma}_n^{-1})^{-1}\right] + o(1), \qquad (59)
\end{aligned}
$$

where $\mathbf{D}_n = (-\mathbf{H}_n + \mathbf{\Sigma}_n^{-1})^{-1}(\mathbf{B}_n + \mathbf{\Sigma}_n^{-1})(-\mathbf{H}_n + \mathbf{\Sigma}_n^{-1})^{-1}$. The details are given in the Supplement.

Combine (53) and (54)-(56), we finally get the desired result.

REFERENCES

AITCHISON, J. (1975): "Goodness of Prediction Fit," *Biometrika*, 62 (3), 547–554. [7, 8, 14]

AITCHISON, J. AND I. R. DUNSMORE (1975): *Statistical Prediction Analysis*, Cambridge University Press, 1 ed. [7]

AKAIKE, H. (1974): "A New Look at the Statistical Model Identification," *IEEE Transactions on Automatic Control*, 19 (6), 716–723. [3, 26]

ALBERT, JAMES H. AND SIDDHARTHA CHIB (1993): "Bayesian Analysis of Binary and Polychotomous Response Data," *Journal of the American Statistical Association*, 88 (422), 669–679. [31]

ANDO, TOMOHIRO AND RUEY TSAY (2010): "Predictive Likelihood for Bayesian Model Selection and Averaging," *International Journal of Forecasting*, 26 (4), 744–763. [11]

ANDREWS, DONALD W. K. (1991): "Heteroskedasticity and Autocorrelation Consistent Covariance Matrix Estimation," *Econometrica*, 59 (3), 817–858. [22]

ANGRIST, JOSHUA D. AND JÖRN-STEFFEN PISCHKE (2009): *Mostly Harmless Econometrics: An Empiricist's Companion*, Princeton, New Jersey Oxford: Princeton University Press. [16]

BARBERIS, NICHOLAS (2000): "Investing for the Long Run When Returns Are Predictable," *The Journal of Finance*, 55 (1), 225–264. [14]

BROWN, LAWRENCE D., EDWARD I. GEORGE, AND XINYI XU (2008): "Admissible Predictive Density Estimation," *The Annals of Statistics*, 36 (3), 1156–1170. [7]

FOURDRINIER, DOMINIQUE, ÉRIC MARCHAND, ALI RIGHI, AND WILLIAM E. STRAWDERMAN (2011): "On Improved Predictive Density Estimation with Parametric Constraints," *Electronic Journal of Statistics*, 5 (none), 172–191. [15]

GEORGE, EDWARD I., FENG LIANG, AND XINYI XU (2006): "Improved Minimax Predictive Densities under Kullback–Leibler Loss," *The Annals of Statistics*, 34 (1), 78–91. [7, 15]

GEORGE, EDWARD I. AND XINYI XU (2008): "Predictive Density Estimation for Multiple Regression," *Econometric Theory*, 24 (02), 528–544. [15]

——— (2010): "Bayesian Predictive Density Estimation," in *Frontiers of Statistical Decision Making and Bayesian Analysis*, Springer, 83–95. [14]

GOOD, I. J. (1952): "Rational Decisions," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 14 (1), 107–114. [8]

GRANGER, CLIVE W.J., MAXWELL L. KING, AND HALBERT WHITE (1995): "Comments on Testing Economic Theories and the Use of Model Selection Criteria," *Journal of Econometrics*, 67 (1), 173–187. [2]

HAMURA, YASUYUKI AND TATSUYA KUBOKAWA (2022): "Bayesian Predictive Density Estimation for a Chi-Squared Model Using Information from a Normal Observation with Unknown Mean and Variance," *Journal of Statistical Planning and Inference*, 217, 33–51. [7]

HANSEN, BRUCE E. (2005): "Challenges for Econometric Model Selection," *Econometric Theory*, 21 (01), 60–68. [2]

HOLMES, CHRIS C. AND LEONHARD HELD (2006): "Bayesian Auxiliary Variable Models for Binary and Multinomial Regression," *Bayesian Analysis*, 1 (1). [31]

KADANE, JOSEPH B AND NICOLE A LAZAR (2004): "Methods and Criteria for Model Selection," *Journal of the American Statistical Association*, 99 (465), 279–290. [2]

KASS, ROBERT E., LUKE TIERNEY, AND JOSEPH B. KADANE (1990): "The Validity of Posterior Expansions Based on Laplace's Method," *Bayesian and Likelihood Methods in Statistics and Econometrics*, 473–488. [38]

KASS, ROBERT E. AND LARRY WASSERMAN (1995): "A Reference Bayesian Test for Nested Hypotheses and Its Relationship to the Schwarz Criterion," *Journal of the American Statistical Association*, 90 (431), 928–934. [27]

KATO, KENGO (2009): "Improved Prediction for a Multivariate Normal Distribution with Unknown Mean and Variance," *Annals of the Institute of Statistical Mathematics*, 61 (3), 531–542. [7, 15]

KIM, SANGJOON, NEIL SHEPHERD, AND SIDDHARTHA CHIB (1998): "Stochastic Volatility: Likelihood Inference and Comparison with ARCH Models," *Review of Economic Studies*, 65 (3), 361–393. [35]

KOMAKI, F (1996): "On Asymptotic Properties of Predictive Distributions," *Biometrika*, 83 (2), 299–313. [14]

KOMAKI, F. (2001): "A Shrinkage Predictive Distribution for Multivariate Normal Observables," *Biometrika*, 88 (3), 859–864. [7, 15]

LEVY, MARTIN S. AND S. K. PERNG (1986): "An Optimal Prediction Function for the Normal Linear Model," *Journal of the American Statistical Association*, 81 (393), 196–198. [14]

LI, YONG, SUSHANTA K. MALLICK, NIANLING WANG, JUN YU, AND TAO ZENG (2025): "Deviance Information Criterion for Bayesian Model Selection: Theoretical Justification and Applications," *Journal of Econometrics*, 105978. [3, 23, 33, 41]

LI, YONG, JUN YU, AND TAO ZENG (2020): "Deviance Information Criterion for Latent Variable Models and Misspecified Models," *Journal of Econometrics*, 216 (2), 450–493. [3, 9, 11, 22, 23, 24, 26]

LIU, JUN S. AND YING NIAN WU (1999): "Parameter Expansion for Data Augmentation," *Journal of the American Statistical Association*, 94 (448), 1264–1274. [31]

MARCHAND, ÉRIC AND ABDOLNASSER SADEGHKHANI (2018): "On Predictive Density Estimation with Additional Information," *Electronic Journal of Statistics*, 12 (2), 4209–4238. [7, 15]

MATSUDA, TAKERU AND FUMIYASU KOMAKI (2015): "Singular Value Shrinkage Priors for Bayesian Prediction," *Biometrika*, 102 (4), 843–854. [15]

MEYER, RENATE AND JUN YU (2000): "BUGS for a Bayesian Analysis of Stochastic Volatility Models," *The Econometrics Journal*, 3 (2), 198–215. [35]

MÜLLER, ULRICH K. (2013): "Risk of Bayesian Inference in Misspecified Models, and the Sandwich Covariance Matrix," *Econometrica*, 81 (5), 1805–1849. [2, 3, 6, 7, 16, 25, 36]

MURRAY, GORDON D. (1977): "A Note on the Estimation of Probability Density Functions," *Biometrika*, 64 (1), 150. [14]

NEWEY, WHITNEY K. AND KENNETH D. WEST (1987): "A Simple, Positive Semi-Definite, Heteroskedasticity and Autocorrelation Consistent Covariance Matrix," *Econometrica*, 55 (3), 703–708. [22]

NG, VEE MING (1980): "On the Estimation of Parametric Density Functions," *Biometrika*, 67 (2), 505–506. [14]

NISHI, KOUHEI, TAKESHI KUROSAWA, AND NOBUYUKI OZEKI (2024): "Dominance of Posterior Predictive Densities over Plug-in Densities for Order Statistics in Exponential Distributions," *Computational Statistics*, 39 (4), 2291–2321. [7, 15]

PHILLIPS, PETER C.B. (1995): "Bayesian Model Selection and Prediction with Empirical Applications," *Journal of Econometrics*, 69 (1), 289–331. [2]

PHILLIPS, PETER C. B. (1996): "Econometric Model Determination," *Econometrica*, 64 (4), 763–812. [2]

POLSON, NICHOLAS G., JAMES G. SCOTT, AND JESSE WINDLE (2013): "Bayesian Inference for Logistic Models Using Pólya–Gamma Latent Variables," *Journal of the American Statistical Association*, 108 (504), 1339–1349. [31]

SPIEGELHALTER, DAVID J., NICOLA G. BEST, BRADLEY P. CARLIN, AND ANGELIKA VAN DER LINDE (2002): "Bayesian Measures of Model Complexity and Fit," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 64 (4), 583–639. [3, 26]

TAKEUCHI, K (1976): "Distribution of information statistics and a criterion of model fitting," *Mathematical Science*, 153, 12–18. [3, 22, 26]

TANNER, MARTIN A. AND WING HUNG WONG (1987): "The Calculation of Posterior Distributions by Data Augmentation," *Journal of the American Statistical Association*, 82 (398), 528–540. [31]

TIERNEY, LUKE, ROBERT E. KASS, AND JOSEPH B. KADANE (1989): "Approximate Marginal Densities of Nonlinear Functions," *Biometrika*, 76 (3), 425–433. [38]

VAN DER VAART, A. W. (1998): *Asymptotic Statistics*, Cambridge University Press, 1 ed. [5]

WHITE, HALBERT (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48 (4), 817–838. [18]

——— (1982): "Maximum Likelihood Estimation of Misspecified Models," *Econometrica*, 50 (1), 1–25. [2, 5]

YU, JUN (2005): "On Leverage in a Stochastic Volatility Model," *Journal of Econometrics*, 127 (2), 165–178. [35]

ZENS, GREGOR, SYLVIA FRÜHWIRTH-SCHNATTER, AND HELGA WAGNER (2023a): "Efficient Bayesian Modeling of Binary and Categorical Data in R: The UPG Package," . [33]

——— (2023b): "Ultimate Pólya Gamma Samplers - Efficient MCMC for Possibly Imbalanced Binary and Categorical Data," . [31]

Co-editor [Name Surname; will be inserted later] handled this manuscript.