Statistica Sinica Preprint No: SS-2015-0380R2	
Title	Information criteria for prediction when distributions of data and target variables are different
Manuscript ID	SS-2015-0380R2
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202015.0380
Complete List of Authors	Keisuke Yano and Fumiyasu Komaki
Corresponding Author	Keisuke Yano
E-mail	Keisuke_Yano@mist.i.u-tokyo.ac.jp

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

INFORMATION CRITERIA FOR PREDICTION WHEN THE DISTRIBUTIONS OF CURRENT AND FUTURE OBSERVATIONS DIFFER

Keisuke Yano[1] and Fumiyasu Komaki[1,2]

[1]*The University of Tokyo and* [2]*RIKEN Brain Science Institute*

Abstract: We consider prediction when distributions of current and future observations may differ. We derive the asymptotic Kullback–Leibler risks of Bayesian predictive distributions when both the numbers of current and future observations grow to infinity. Based on these results, we construct model selection criteria when the true distributions of current and future observations may differ. Through numerical experiments, we show that Bayesian predictive distributions based on the proposed model selection criteria work well.

Key words and phrases: Bayesian prediction, Kullback–Leibler divergence, Curve extrapolation, Model selection, Bootstrap

1. Introduction

We consider prediction of future observations on the basis of current observations. Let $x^{(N)} = (x_1, \ldots, x_N)^\top$ be current observations and let

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

$y^{(M)} = (x_{N+1}, \ldots, x_{N+M})^\top$ be future observations. The joint distribution of $x^{(N)}$ and $y^{(M)}$ is assumed to be a distribution $p_{\theta^*}(x^{(N)}) \times q_{\theta^*}(y^{(M)})$ in a parametric model $\{p_\theta(x^{(N)}) \times q_\theta(y^{(M)}) : \theta \in \Theta\}$, where Θ is a parameter space in $\mathbb{R}^d$. Here we assume that distributions $p_\theta(x^{(N)})$ and $q_\theta(y^{(M)})$ may differ with an unknown parameter θ. In this paper, we consider all random variables $x_1, \ldots, x_N, x_{N+1}, \ldots, x_{N+M}$ to be independent in the model and also consider that the dimension d is fixed and does not grow to infinity as the number N grows to infinity.

In applications, we often encounter parametric models in which distributions $p_\theta(x^{(N)})$ and $q_\theta(y^{(M)})$ differ. The widely used linear regression model is a typical example. Suppose that the distribution $p_\theta(x^{(N)})$ is given by the N-dimensional Gaussian distribution $\mathcal{N}_N(Z\theta, \sigma^2 I_N)$, where Z is an $N \times d$-dimensional design matrix and θ is a d-dimensional parameter, and suppose that the distribution $q_\theta(y^{(M)})$ is given by the M-dimensional Gaussian distribution $\mathcal{N}_M(\tilde{Z}\theta, \sigma^2 I_M)$, where $\tilde{Z}$ is an $M \times d$-dimensional design matrix. For example, for an unknown function $f(t)$ and a random function $\varepsilon(t)$ of which each distribution at t is independent and $\mathcal{N}_1(0, \sigma^2)$ with known $\sigma^2 > 0$, we observe $x^{(N)} = (x(t_1), x(t_2), \ldots, x(t_N))^\top$ from a curve $x(t)$ given by $x(t) = f(t) + \varepsilon(t)$ at $(t_1, \ldots, t_N)$, and predict $y^{(M)} = (x(t_{N+1}), x(t_{N+2}), \ldots, x(t_{N+M}))^\top$ from the curve $x(t)$ at $(t_{N+1}, \ldots, t_{N+M})$. For simplicity, we model $f(t)$ as $f(t) = \theta_1 + \theta_2 t$. Thus, the parametric model for $x^{(N)}$ and $y^{(M)}$ is given by $\{\phi_N(x^{(N)}; Z\theta, \sigma^2 I_N) \times \phi_M(y^{(M)}; \tilde{Z}\theta, \sigma^2 I_M) :$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

$\theta \in \mathbb{R}^2\}$ where $\theta = (\theta_1, \theta_2)^\top$, $\phi_d(\cdot; \mu, \Sigma)$ is the density of $\mathcal{N}_d(\mu, \Sigma)$,

$$Z = \begin{pmatrix} 1 & t_1 \\ \dots & \dots \\ 1 & t_N \end{pmatrix} \text{ and } \tilde{Z} = \begin{pmatrix} 1 & t_{N+1} \\ \dots & \dots \\ 1 & t_{N+M} \end{pmatrix}.$$

We discuss the details of the above example in Section 4.

First, we discuss performance of a predictive distribution $\hat{q}(y^{(M)}; x^{(N)})$ when the true distributions $p_{\theta^*}(x^{(N)})$ and $q_{\theta^*}(y^{(M)})$ may differ. We measure the performance of $\hat{q}(y^{(M)}; x^{(N)})$ using the Kullback–Leibler risk

$$\int p_{\theta^*}(x^{(N)}) \int q_{\theta^*}(y^{(M)}) \log \frac{q_{\theta^*}(y^{(M)})}{\hat{q}(y^{(M)}; x^{(N)})} \mathrm{d}y^{(M)} \mathrm{d}x^{(N)}.$$

We show that the Bayesian predictive distribution based on prior density $\pi(\theta)$ on Θ,

$$q_\pi(y^{(M)}|x^{(N)}) := \frac{\int q_\theta(y^{(M)}) p_\theta(x^{(N)}) \pi(\theta) \mathrm{d}\theta}{\int p_\theta(x^{(N)}) \pi(\theta) \mathrm{d}\theta},$$

has a smaller Kullback–Leibler risk than the plug-in predictive distribution based on the maximum likelihood estimator $\hat{\theta}$ when both the numbers N and M grow to infinity. Komaki (1996, 2015) and Hartigan (1998) showed that the Bayesian predictive distribution has a smaller Kullback–Leibler risk than the plug-in predictive distribution when M is 1. Our result is an extension of these works to settings in which both the numbers N and M grow to infinity, which often appear. The typical example is the example in the previous paragraph.

Second, we discuss model selection when the true distributions of $x^{(N)}$ and $y^{(M)}$ may differ. For such a setting, we propose model selection criteria that are asymptotically unbiased estimators of the Kullback–Leibler

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

risk of the Bayesian predictive distribution. Akaike's Information Criterion (AIC; Akaike (1973)) and Predictive Information Criterion (PIC; Kitagawa (1997)) constitute such estimators for the plug-in predictive distribution and the Bayesian predictive distribution, respectively, when the true distribution of $y^{(M)}$ is identical to that of $x^{(N)}$. However, these criteria are not asymptotically unbiased when the true distributions of $x^{(N)}$ and $y^{(M)}$ differ; our model selection criteria represent the extensions of PIC to this setting. We extend PIC instead of AIC because the Bayesian predictive distribution has a smaller Kullback–Leibler risk than the plug-in predictive distribution.

This paper is organized as follows. In Section 2, we discuss the performance of the Bayesian predictive distribution. In Section 3, we propose a model selection criterion and its bootstrap adjustment. In Section 4, we present numerical experiments to compare performances of several model selection criteria.

2. Bayesian predictive distributions when both the numbers N and M grow to infinity

In this section, we discuss the Kullback–Leibler risk of the predictive distribution. We consider two parametric models for the joint distribution of $x^{(N)}$ and $y^{(M)}$, the full model $\mathcal{M} := \{p_\theta(x^{(N)}) \times q_\theta(y^{(M)}) : \theta \in \Theta \subset \mathbb{R}^d\}$ and the sub-model $\mathcal{M}_{\rm s} := \{p_{\theta_{\rm s}}(x^{(N)}) \times q_{\theta_{\rm s}}(y^{(M)}) : \theta_{\rm s} \in \Theta_{\rm s} \subset \Theta\}$, where we denote the dimension of $\Theta_{\rm s}$ by $d_{\rm s}\,(\leq d)$. We discuss the Kullback–Leibler risk of the predictive distribution based on the sub-model. Since the sub-model is included in the full model, we decompose parameter θ in Θ into

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

$\theta(\theta_{\rm s}, \gamma_{\rm s})$ using an additional parameter $\gamma_{\rm s} \in \mathbb{R}^{d-d_{\rm s}}$, where $\theta(\theta_{\rm s}, 0)$ takes a value in $\Theta_{\rm s}$. We denote the true parameter value with respect to $(\theta_{\rm s}, \gamma_{\rm s})$ by $(\theta_{\rm s}^*, \gamma_{\rm s}^*)$.

We assume that the true parameter value is given by

$$\begin{pmatrix} \theta_{\rm s}^* \\ \gamma_{\rm s}^* \end{pmatrix} = \begin{pmatrix} \theta_{\rm s,0} \\ 0 \end{pmatrix} + \frac{1}{\sqrt{N}} h_{\rm s} \tag{2.1}$$

for some $\theta_{\rm s,0} \in \Theta_{\rm s}$ and some $h_{\rm s} \in \mathbb{R}^d$. This assumption is called local misspecification and is an extension of the setting in which the true distribution is included in the sub-model. For further discussions concerning local misspecification, see Shimodaira (1997), Hjort and Claeskens (2003), and Claeskens and Hjort (2003).

We denote the Kullback–Leibler risk of predictive distribution $\hat{q}(y^{(M)}; x^{(N)})$ by

$$R_{\rm s}(\theta^*, \hat{q}) = \int p_{\theta^*}(x^{(N)}) \int q_{\theta^*}(y^{(M)}) \log \frac{q_{\theta^*}(y^{(M)})}{\hat{q}(y^{(M)}; x^{(N)})} {\rm d}y^{(M)} {\rm d}x^{(N)}$$

and denote the Bayesian predictive distribution based on prior density $\pi_{\rm s}(\theta_{\rm s})$ on $\Theta_{\rm s}$ by

$$q_{\pi_{\rm s}}(y^{(M)}|x^{(N)}) = \frac{\int q_{\theta_{\rm s}}(y^{(M)}) p_{\theta_{\rm s}}(x^{(N)}) \pi_{\rm s}(\theta_{\rm s}) {\rm d}\theta_{\rm s}}{\int p_{\theta_{\rm s}}(x^{(N)}) \pi_{\rm s}(\theta_{\rm s}) {\rm d}\theta_{\rm s}}.$$

The Fisher information matrices of $p_{\theta_{\rm s}}(x^{(N)})$ and $q_{\theta_{\rm s}}(y^{(M)})$ at $\theta_{\rm s,0}$ are $g_{\rm s}^{(N)}(\theta_{\rm s,0})$ and $\tilde{g}_{\rm s}^{(M)}(\theta_{\rm s,0})$, respectively, and the Fisher information matrix of $q_\theta(y^{(M)})$ with respect to $(\theta_{\rm s}, \gamma_{\rm s})$ at $(\theta_{\rm s}^*, \gamma_{\rm s}^*)$ is $\tilde{g}^{(M)}(\theta_{\rm s}^*, \gamma_{\rm s}^*)$. For $a \in \mathbb{N}$ and $b \in \mathbb{N}$, we denote the $a \times b$-dimensional zero matrix by $O_{a,b}$.

Theorem 1. *If M is given by a constant multiple of N and (2.1) holds, as N grows to infinity, the Kullback–Leibler risk of the Bayesian predictive*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

distribution based on prior density π_{s} *on* Θ_{s} *is*

$$R_{\mathrm{s}}(\theta^*, q_{\pi_{\mathrm{s}}}) = \frac{1}{2N} h_{\mathrm{s}}^\top G_{\mathrm{s}}^{(N)}(\theta^*, \theta_{\mathrm{s},0}) h_{\mathrm{s}} + \frac{1}{2} \log \frac{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0}) + \tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})|}{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})|} + \mathrm{o}(1), \tag{2.2}$$

where $|\cdot|$ *is the determinant and*

$$G_{\mathrm{s}}^{(N)}(\theta^*, \theta_{\mathrm{s},0}) := \left(\tilde{g}^{(M)^{-1}}(\theta_{\mathrm{s}}^*, \gamma_{\mathrm{s}}^*) + \begin{pmatrix} g_{\mathrm{s}}^{(N)^{-1}}(\theta_{\mathrm{s},0}) & O_{d_{\mathrm{s}},(d-d_{\mathrm{s}})} \\ O_{(d-d_{\mathrm{s}}),d_{\mathrm{s}}} & O_{(d-d_{\mathrm{s}}),(d-d_{\mathrm{s}})} \end{pmatrix} \right)^{-1}.$$

The proof is given in the Appendix.

Remark 1. Consider invariance and the dependence on prior densities. First, from (29) in the Appendix, expansion (2.2) is invariant up to o(1) under the reparameterization of θ. Second, the asymptotic Kullback–Leibler risk of the Bayesian predictive distribution does not depend on prior densities up to o(1), which corresponds to the fact that the N^{-1}-order term of the asymptotic Kullback–Leibler risk when $M = 1$ does not depend on prior densities, as discussed in Remark 2 below.

Remark 2. The asymptotic Kullback–Leibler risk (2.2) is considered to be the accumulation of the asymptotic Kullback–Leibler risks of the Bayesian predictive distribution when M is 1. We explain this heuristically. For simplicity, we assume that h_{s} vanishes and that all random variables $\{x_i : i = 1, \ldots, N + M\}$ are identically distributed. Then, the constant order term in (2.2) is given by $(d_{\mathrm{s}}/2) \log\{(N + M)/N\}$, whereas the N^{-1}-order term of the Kullback–Leibler risk when $M = 1$ is given by $(d_{\mathrm{s}}/2N)$. Since the Bayesian predictive distribution $q_{\pi_{\mathrm{s}}}(y^{(M)}|x^{(N)})$ is decomposed into a

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

product of Bayesian predictive distributions as

$$q_{\pi_{\rm s}}(y^{(M)}|x^{(N)}) = q_{\pi_{\rm s}}(x_{N+1}|x^{(N)}) \times q_{\pi_{\rm s}}(x_{N+2}|x^{(N)}, y^{(1)}) \times \cdots \times q_{\pi_{\rm s}}(x_{N+M}|x^{(N)}, y^{(M-1)}), \tag{2.3}$$

the Kullback–Leibler risk $R_{\rm s}(\theta^*, q_{\pi_{\rm s}})$ is also calculated as $\sum_{j=1}^{M}\{d_{\rm s}/(2N + 2j)\}$. This is asymptotically equal to $(d_{\rm s}/2)\log\{(N+M)/N\}$.

Remark 3. Consider the predictive setting in which M is fixed and N grows to infinity. In this setting, from Lemmas A1 and A2 and (19) in the Appendix, the Kullback–Leibler risk of the Bayesian predictive distribution is given by

$$R_{\rm s}(\theta^*, q_{\pi_{\rm s}}) = \frac{1}{2N} h_{\rm s}^{\top} \tilde{g}^{(M)}(\theta^*) h_{\rm s} + \frac{1}{2N} {\rm tr}\{g_{\rm s}^{(N)-1}(\theta_{{\rm s},0}) \tilde{g}_{\rm s}^{(M)}(\theta_{{\rm s},0})\} + {\rm o}(1/N).$$

This can also be derived by further expanding the right-hand side of (2.2) with respect to N. By setting M equal to 1 and allowing $h_{\rm s}$ to vanish, this is consistent with the result in Komaki (1996, 2015) and Hartigan (1998).

The following theorem states that the Bayesian predictive distribution is asymptotically better than the plug-in predictive distribution based on the maximum likelihood estimator when both the numbers N and M simultaneously grow to infinity. The plug-in predictive distribution based on the maximum likelihood estimator $\hat{\theta}_{\rm s}$ is

$$q(y^{(M)}; \hat{\theta}_{\rm s}(x^{(N)})) = q_{\theta_{\rm s}=\hat{\theta}_{\rm s}(x^{(N)})}(y^{(M)}).$$

Theorem 2. *If M is given by a constant multiple of N and (2.1) holds, as N grows to infinity, the difference between the Kullback–Leibler risks of the*

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Bayesian predictive distribution based on prior density $\pi_{\mathrm{s}}(\theta_{\mathrm{s}})$ *on* Θ_{s} *and the plug-in predictive distribution with the maximum likelihood estimator* $\hat{\theta}_{\mathrm{s}}$ *is*

$$\begin{aligned} &R_{\mathrm{s}}(\theta^*, q_{\pi_{\mathrm{s}}}) - R_{\mathrm{s}}(\theta^*, q(\cdot; \hat{\theta}_{\mathrm{s}})) \\ &\quad = \frac{1}{2}\left[\log \frac{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0}) + \tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})|}{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})|} - \mathrm{tr}\{{g_{\mathrm{s}}^{(N)}}^{-1}(\theta_{\mathrm{s},0})\tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})\}\right] \\ &\qquad + \frac{1}{2}\left\{h_{\mathrm{s}}^{\top} G_{\mathrm{s}}^{(N)}(\theta^*, \theta_{\mathrm{s},0}) h_{\mathrm{s}} - h_{\mathrm{s}}^{\top}\tilde{g}^{(M)}(\theta_{\mathrm{s}}^*, \gamma_{\mathrm{s}}^*) h_{\mathrm{s}}\right\} + \mathrm{o}(1), \end{aligned} \tag{2.4}$$

where tr *is the trace of a matrix. Furthermore, the difference up to* $\mathrm{o}(1)$ *is non-positive. If two Fisher information matrices* $g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})$ *and* $\tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})$ *are positive definite, the difference up to* $\mathrm{o}(1)$ *is strictly negative.*

Proof. From (18) in the Appendix, the Kullback–Leibler risk $R_{\mathrm{s}}(\theta^*, q(\cdot; \hat{\theta}_{\mathrm{s}}))$ is expanded as

$$R_{\mathrm{s}}(\theta^*, q(\cdot; \hat{\theta}_{\mathrm{s}})) = \frac{1}{2N} h_{\mathrm{s}}^{\top}\tilde{g}^{(M)}(\theta_{\mathrm{s}}^*, \gamma_{\mathrm{s}}^*) h_{\mathrm{s}} + \frac{1}{2}\mathrm{tr}\{{g_{\mathrm{s}}^{(N)}}^{-1}(\theta_{\mathrm{s},0})\tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})\} + \mathrm{o}(1).$$

By combining this with Theorem 1, we obtain (2.4).

Since $\log|I + A^{-1}B| \leq \mathrm{tr}\{A^{-1}B\}$ for the invertible matrix A and the matrix B, we have

$$\log \frac{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0}) + \tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})|}{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})|} \leq \mathrm{tr}\{{g_{\mathrm{s}}^{(N)}}^{-1}(\theta_{\mathrm{s},0})\tilde{g}_{\mathrm{s}}^{(M)}(\theta_{\mathrm{s},0})\}.$$

If the two matrices are positive definite, the strict inequality holds. Since $A - (A^{-1} + B)^{-1}$ is positive semidefinite for the invertible matrix A and the matrix B, we have $h_{\mathrm{s}}^{\top}\tilde{g}^{(M)}(\theta_{\mathrm{s}}^*, \gamma_{\mathrm{s}}^*) h_{\mathrm{s}} \geq h_{\mathrm{s}}^{\top} G_{\mathrm{s}}^{(N)}(\theta^*, \theta_{\mathrm{s},0}) h_{\mathrm{s}}$. Thus, we complete the proof. □

Remark 4. The implication of Theorem 2 is explained by Remark 2: the one-step ahead Bayesian predictive distribution $q_{\pi_{\mathrm{s}}}(x_{N+i}|x^{(N)}, y^{(i-1)})$ is up-

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

dated based on $y^{(i-1)}$ as (2.3), whereas the one-step ahead plug-in predictive distribution $q(x_{N+i}; \hat{\theta}(x^{(N)}))$ is not.

Remark 5. The result is related to prediction in locally asymptotically mixed normal (LAMN) models as discussed in Sei and Komaki (2007). In both our setting and LAMN models, we consider prediction of future observations based on the current observations conditioned on two Fisher information matrices of current and future observations. Indeed, the Kullback–Leibler risk (2.2) has the same form as (2) in Sei and Komaki (2007).

3. Information criteria when the true distributions of current and future observations differ

First, to construct information criteria, we derive an asymptotically unbiased estimator of the Kullback–Leibler risk of the Bayesian predictive distribution when the true distributions of current and future observations differ. Let $\hat{\theta}_{\mathrm{s}}$ be the maximum likelihood estimator of θ_{s} and let $\hat{\xi}_{\mathrm{s}}$ be the maximum likelihood estimator of $(\theta_{\mathrm{s}}^{\top}, \gamma_{\mathrm{s}}^{\top})^{\top}$, respectively. Let $\hat{h}_{\mathrm{s}}$ be given by

$$\hat{h}_{\mathrm{s}} := \sqrt{N} \left\{ \hat{\xi}_{\mathrm{s}} - \begin{pmatrix} \hat{\theta}_{\mathrm{s}} \\ 0 \end{pmatrix} \right\}.$$

Let $\hat{G}_{\mathrm{s}}^{(N)}$ be a matrix given by

$$\hat{G}_{\mathrm{s}}^{(N)} := \left(\tilde{g}^{(M)^{-1}}(\hat{\xi}_{\mathrm{s}}) + \begin{pmatrix} g_{\mathrm{s}}^{(N)^{-1}}(\hat{\theta}_{\mathrm{s}}) & O_{d_{\mathrm{s}},(d-d_{\mathrm{s}})} \\ O_{(d-d_{\mathrm{s}}),d_{\mathrm{s}}} & O_{(d-d_{\mathrm{s}}),(d-d_{\mathrm{s}})} \end{pmatrix} \right)^{-1}$$

and let $\hat{G}_{\mathrm{s},d_{\mathrm{s}}\times d_{\mathrm{s}}}^{(N)}$ be the top left $d_{\mathrm{s}} \times d_{\mathrm{s}}$-dimensional sub-matrix of $\hat{G}_{\mathrm{s}}^{(N)}$.

Let $\hat{R}_{\mathrm{s}}$ be the estimator of the Kullback–Leibler risk of the Bayesian

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

predictive distribution based on sub-model $\mathcal{M}_{\rm s}$ given by

$$\hat{R}_{\rm s} := \frac{1}{2N}\hat{h}_{\rm s}^\top \hat{G}_{\rm s}^{(N)}\hat{h}_{\rm s} + \frac{1}{2}{\rm tr}\{{g_{\rm s}^{(N)}}^{-1}(\hat{\theta}_{\rm s})\hat{G}^{(N)}_{{\rm s},d_{\rm s}\times d_{\rm s}}\} - \frac{1}{2}{\rm tr}\{{g^{(N)}}^{-1}(\hat{\xi}_{\rm s})\hat{G}_{\rm s}^{(N)}\} + \frac{1}{2}\log\frac{|g_{\rm s}^{(N)}(\hat{\theta}_{\rm s}) + \tilde{g}_{\rm s}^{(M)}(\hat{\theta}_{\rm s})|}{|g_{\rm s}^{(N)}(\hat{\theta}_{\rm s})|}. \tag{3.1}$$

Theorem 3. *If M is given by a constant multiple of N and (2.1) holds, for prior density $\pi_{\rm s}(\theta_{\rm s})$ on $\Theta_{\rm s}$, $\hat{R}_{\rm s}$ given as (3.1) is an asymptotically unbiased estimator of the Kullback–Leibler risk $R_{\rm s}(\theta^*, q_{\pi_{\rm s}})$ of the Bayesian predictive distribution $q_{\pi_{\rm s}}$.*

The proof is in the Appendix.

Remark 6. Since estimator $(1/N)\hat{h}_{\rm s}^\top \hat{G}_{\rm s}^{(N)}\hat{h}_{\rm s}$ of $(1/N)h_{\rm s}^\top G^{(N)}(\theta^*, \theta_{{\rm s},0})h_{\rm s}$ has an asymptotic bias

$${\rm tr}\{{g^{(N)}}^{-1}(\theta^*)G_{\rm s}^{(N)}(\theta^*, \theta_{{\rm s},0})\} - {\rm tr}\{{g_{\rm s}^{(N)}}^{-1}(\theta_{{\rm s},0})G^{(N)}_{{\rm s},d_{\rm s}\times d_{\rm s}}(\theta^*, \theta_{{\rm s},0})\}$$

as shown in (33) in the Appendix, the bias adjustment term

$${\rm tr}\{{g_{\rm s}^{(N)}}^{-1}(\hat{\theta}_{\rm s})\hat{G}^{(N)}_{{\rm s},d_{\rm s}\times d_{\rm s}}\} - {\rm tr}\{{g^{(N)}}^{-1}(\hat{\xi}_{\rm s})\hat{G}_{\rm s}^{(N)}\}$$

appears in (3.1).

On the basis of Theorem 3, we propose an information criterion when the true distributions of current and future observations may differ. We prepare $K+1$ parametric models for the joint distribution of $x^{(N)}$ and $y^{(M)}$, the full model $\mathcal{M}$ and the K sub-models $\{\mathcal{M}_m\}_{m=1}^K$, where each sub-model $\mathcal{M}_m$ is included in $\mathcal{M}$. We denote the dimension of sub-model $\mathcal{M}_m$ by d_m for $m \in \{1, \ldots, K\}$ and the dimension of the full model by d_{K+1}, for convenience.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

As an information criterion, we propose a multistep predictive information criterion (MSPIC) given by

$$\begin{aligned}\mathrm{MSPIC}(m) := \frac{1}{N}\hat{h}_m^\top \hat{G}_m^{(N)}\hat{h}_m + \mathrm{tr}\{{g_m^{(N)}}^{-1}(\hat{\theta}_m)\hat{G}^{(N)}_{m,d_m\times d_m}\} \\ - \mathrm{tr}\{{g^{(N)}}^{-1}(\hat{\xi}_m)\hat{G}_m^{(N)}\} + \log\frac{|g_m^{(N)}(\hat{\theta}_m)+\tilde{g}_m^{(M)}(\hat{\theta}_m)|}{|g_m^{(N)}(\hat{\theta}_m)|},\end{aligned}$$

where the quantities with subscript m are those with subscript s when the sub-model $\mathcal{M}_{\mathrm{s}}$ is $\mathcal{M}_m$. If the true distribution satisfies (2.1) for a sub-model $\mathcal{M}_m$, MSPIC(m) is an asymptotically unbiased estimator of $2\times R_m(\theta^*, q_{\pi_m})$ according to Theorem 3. We name the model selection criterion MSPIC because size M plays the role of time as discussed in Remark 2. If two Fisher information matrices $g_m^{(N)}(\theta_m)$ and $\tilde{g}_m^{(M)}(\theta_m)$ are identical, MSPIC(m) coincides with PIC (Kitagawa (1997)) when using the uniform prior and also with the predictive likelihood (Akaike (1980)).

MSPIC itself is an estimator of the risk and may have an excessive variance and an excessive skewness. These excesses may appear in the first three terms in the definition of MSPIC because matrix $\hat{G}_m^{(N)}$ is not equal to the asymptotic covariance matrix of $\hat{h}_m$.

To reduce the effects of the variance and the skewness of MSPIC, we use its bootstrap adjustment $\mathrm{MSPIC}_{\mathrm{BS}}$. First, we generate B bootstrap samples $x_1^{(N)},\ldots,x_b^{(N)},\ldots,x_B^{(N)}$ according to the bootstrap method using the full model. Second, for each $b\in\{1,\ldots,B\}$, we calculate the value of $\mathrm{MSPIC}_1(m;x_b^{(N)})$ defined by

$$\begin{aligned}\mathrm{MSPIC}_1(m;x_b^{(N)}) := \frac{1}{N}\hat{h}_m^\top\hat{G}^{(N)}\hat{h}_m \\ + \mathrm{tr}\{\hat{G}^{(N)}_{m,d_m\times d_m}{g_m^{(N)}}^{-1}(\hat{\theta}_m)\} - \mathrm{tr}\{\hat{G}_m^{(N)}{g^{(N)}}^{-1}(\hat{\xi}_{\mathrm{m}})\}\end{aligned}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

using $x_b^{(N)}$ instead of $x^{(N)}$. Finally, we obtain

$$\mathrm{MSPIC}_{\mathrm{BS}}(m) := \frac{1}{B}\Sigma_{b=1}^{B}\mathrm{MSPIC}_1(m; x_b^{(N)}) + \log \frac{|g_m^{(N)}(\hat{\theta}_m) + \tilde{g}_m^{(M)}(\hat{\theta}_m)|}{|g_m^{(N)}(\hat{\theta}_m)|}.$$

Our bootstrap adjustment $\mathrm{MSPIS}_{\mathrm{BS}}$ is related to the bootstrap adjustment of Takeuchi's Information Criterion (TIC; Takeuchi (1976)) considered in Lv and Liu (2014), because TIC also may have an excessive variance and an excessive skewness.

4. Numerical experiments

Through two numerical experiments, we show that the proposed model selection criteria can effectively perform predictions when the distributions of current and future observations differ.

For comparison, we used AIC, PIC, the bootstrap adjustment of PIC ($\mathrm{PIC}_{\mathrm{BS}}$) , MSPIC, and $\mathrm{MSPIC}_{\mathrm{BS}}$ and evaluated their performances based on the goodness of the Bayesian predictive distributions using their selected models. We considered the goodness of predictive distributions as follows. We generated current and future observations R times and calculated the mean of negative log Bayesian predictive densities, $-\Sigma_{r=1}^{R}\log q_\pi(y_r^{(M)}|x_r^{(N)})/R$, based on the selected model by each criterion. Here, for $r = 1, \ldots, R$, $x_r^{(N)}$ and $y_r^{(M)}$ were the r-th current observations and the r-th future observations, respectively. It is desirable that the value be small because it is an estimator of the Kullback–Leibler risk up to the term independent of the predictive distribution. We set $R = 100$ in the first experiment and $R = 50$ in the second.

4.1. The extrapolation of curve fitting

The following setting deals with the high-dimensional model selection.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Table 4.1: The means (with the standard deviations) of negative log predictive densities when the true function is f_1 given in (4.1) and α is 1. The lowest value in each row is underlined.

N and M	AIC	PIC	PIC_{BS}	MSPIC	MSPIC_{BS}
100 and 100	−9.17 (7.92)	−9.34 (7.82)	<u>−9.60</u> (7.15)	−9.34 (7.82)	<u>−9.60</u> (7.16)
100 and 200	−21.0 (12.2)	−21.0 (12.1)	<u>−22.9</u> (10.2)	−21.4 (11.3)	−22.7 (10.1)
100 and 500	−63.3 (23.9)	−63.4 (22.1)	−69.3 (17.5)	−68.3 (18.7)	<u>−69.6</u> (16.5)
100 and 1000	−139 (41.2)	−141 (39.4)	<u>−156</u> (24.4)	−154 (26.8)	<u>−156</u> (21.8)

We do not provide their theoretical extensions. Those are important. We provide the numerical experiment related to the extension.

We consider an extrapolation of the curve fitting mentioned in the introduction. Suppose that a curve $x(t)$ is given by $x(t) = f(t) + \varepsilon(t)$ where $f(t)$ is an unknown function and $\varepsilon(t)$ is a random function. Suppose that we observe $x^{(N)} = (x(t_1), \ldots, x(t_N))^\top$ of $x(t)$ at $(t_1, \ldots, t_N)$ and predict $y^{(M)} = (x(t_{N+1}), \ldots, x(t_{N+M}))^\top$ of $x(t)$ at $(t_{N+1}, \ldots, t_{N+M})$.

In this experiment, we set $t_i = \alpha \times (i/N)$ with $\alpha \in [0, 1]$ for each $i \in \{1, \ldots, N+M\}$. We took $\varepsilon^{(N)} = (\varepsilon(t_1), \ldots, \varepsilon(t_N))^\top$ distributed according to $\mathcal{N}_N(0, \sigma^2 I_N)$ with known σ^2 and $\tilde{\varepsilon}^{(M)} = (\varepsilon(t_{N+1}), \ldots, \varepsilon(t_{N+M}))^\top$ distributed according to $\mathcal{N}_M(0, \sigma^2 I_M)$ with known σ^2. For simplicity, we assumed that $\varepsilon^{(N)}$ and $\tilde{\varepsilon}^{(M)}$ were independent.

We used the following regression models. For each $m \in \{1, \ldots, K+1\}$, the model $\mathcal{M}_m$ is given by $\{\phi_N(x^{(N)}; Z_m\theta_m, \sigma^2 I_N) \times \phi_M(y^{(M)}; \tilde{Z}_m\theta_m, \sigma^2 I_M) : \theta_m \in \mathbb{R}^m\}$, where $\phi_d(\cdot; \mu, \Sigma)$ is the density of $\mathcal{N}_d(\mu, \Sigma)$, and Z_m and $\tilde{Z}_m$ are

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Table 4.2: The means (with the standard deviations) of negative log predictive densities when the true function is f_2 given in (4.2) and α is 1. The lowest value in each row is underlined.

N and M	AIC	PIC	PIC_{BS}	MSPIC	MSPIC_{BS}
100 and 100	−12.8 (7.51)	<u>−13.0</u> (7.32)	<u>−13.0</u> (7.17)	<u>−13.0</u> (7.32)	<u>−13.0</u> (7.17)
100 and 200	−27.6 (10.2)	−28.0 (10.1)	−28.1 (10.1)	<u>−28.4</u> (10.1)	−28.2 (10.1)
100 and 500	−71.5 (18.8)	−72.7 (18.9)	−76.1 (16.6)	−73.8 (19.0)	<u>−76.6</u> (16.4)
100 and 1000	−158 (26.1)	−158 (26.4)	−166 (23.0)	−163 (23.7)	<u>−167</u> (23.3)

design matrices defined by

$$Z_m = \begin{pmatrix} \psi_1(t_1) & \dots & \psi_m(t_1) \\ \dots & \dots & \dots \\ \psi_1(t_N) & \dots & \psi_m(t_N) \end{pmatrix} \quad \text{and} \quad \tilde{Z}_m = \begin{pmatrix} \psi_1(t_{N+1}) & \dots & \psi_m(t_{N+1}) \\ \dots & \dots & \dots \\ \psi_1(t_{N+M}) & \dots & \psi_m(t_{N+M}) \end{pmatrix},$$

respectively. Here $\{\psi_1, \dots, \psi_{K+1}\}$ is a set of functions of t. For $\{\psi_m\}_{m=1}^{K+1}$, we used trigonometric functions $\{\psi_{\text{tri},m}\}_{m=1}^{K+1}$:

$$\psi_{\text{tri},m}(t) = \begin{cases} 1 & (m = 1), \\ \sqrt{2}\cos(2\pi\frac{m}{2}t) & (m : \text{even}), \\ \sqrt{2}\sin(2\pi\frac{m-1}{2}t) & (m : \text{odd}). \end{cases}$$

We compared the negative log Bayesian predictive densities based on the uniform prior for the AIC-best, PIC-best, the PIC_{BS}-best, the MSPIC-best, and the MSPIC_{BS}-best models. For each $m \in \{1, \dots, K+1\}$, the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Table 4.3: The means (with the standard deviations) of negative log predictive densities when the true function is f_2 given as (4.2) and α is 0.9. The lowest value in each row is underlined.

N and M	AIC	PIC	PIC_{BS}	MSPIC	MSPIC_{BS}
100 and 100	−12.5 (8.24)	−12.6 (7.92)	−12.2 (7.89)	<u>−12.7</u> (8.02)	−12.2 (7.80)
100 and 200	−27.3 (12.0)	−27.3 (11.6)	−28.0 (11.3)	−27.4 (11.5)	<u>−28.1</u> (11.2)
100 and 500	−70.8 (19.7)	−71.6 (19.6)	−76.2 (16.9)	−73.3 (20.0)	<u>−76.6</u> (17.1)
100 and 1000	−153 (27.8)	−153 (27.8)	−161 (25.1)	−155 (29.8)	<u>−164</u> (23.2)

negative log Bayesian predictive density based on model $\mathcal{M}_m$ was

$$
\begin{aligned}
-\log q_{u_m}(y^{(M)}|x^{(N)}) = \frac{1}{2\sigma^2}\left|\begin{pmatrix} x^{(N)} \\ y^{(M)} \end{pmatrix} - \begin{pmatrix} Z_m \\ \tilde{Z}_m \end{pmatrix}\hat{\theta}_m(x^{(N)}, y^{(M)})\right|^2 \\
- \frac{1}{2\sigma^2}\left|{x^{(N)}}^\top - Z_m\hat{\theta}_m(x^{(N)})\right|^2 \\
+ \frac{M}{2}\log(2\pi\sigma^2) + \frac{1}{2}\log\frac{|Z_m^\top Z_m + \tilde{Z}_m^\top \tilde{Z}_m|}{|Z_m^\top Z_m|},
\end{aligned}
$$

where we denote the uniform prior on $\mathbb{R}^m$ by u_m and the maximum likelihood estimator of θ_m based on $x^{(N)}$ and $y^{(M)}$ by $\hat{\theta}_m(x^{(N)}, y^{(M)})$.

First, we considered that true function f was

$$
\begin{aligned}
f_1(t) &= 2\sin(2\pi \times t) + 0.2\sin(2\pi \times 4t) \\
&\quad +0.1\sin(2\pi \times 8t) + 0.1\sin(2\pi \times 12t). \qquad (4.1)
\end{aligned}
$$

We set $\sigma^2 = (0.2)^2$, $\alpha = 1.0$, and $K = 30$. Table 4.1 shows that MSPIC_{BS} and PIC_{BS} perform well.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Second, we considered the true function f to be given by

$$f_2(t) = \frac{\pi^2}{6} - \frac{\pi}{2}(t \bmod 2\pi) + \frac{1}{4}(t \bmod 2\pi)^2. \tag{4.2}$$

We set $\sigma^2 = (0.2)^2$ and $K = 15$, and considered the settings with $\alpha = 1$ and with $\alpha = 0.9$. Tables 4.2 and 4.3 show that $\mathrm{MSPIC}_{\mathrm{BS}}$ tends to perform well.

In the second setting, we consider the misspecification discussed in Takeuchi (1976), Sin and White (1996), Fushiki (2005), and Lv and Liu (2014): the true function f_2 is not included in the full model. The experiment indicates that $\mathrm{MSPIC}_{\mathrm{BS}}$ works well in this setting and that the dominance of $\mathrm{MSPIC}_{\mathrm{BS}}$ is enlarged as the ratio of N and M grows.

4.2. The linear regression with an unknown variance

We considered a linear regression with an unknown variance. We assumed that the true distributions of $x^{(N)}$ and $y^{(M)}$ were given by $\mathcal{N}_N(Z\theta^*, \sigma^{*2} I_N)$ and $\mathcal{N}_M(\tilde{Z}\theta^*, \sigma^{*2} I_M)$, respectively. Here Z and $\tilde{Z}$ were a known $N \times 9$-dimensional matrix and a known $M \times 9$-dimensional matrix, respectively. We set $\theta^* = (-0.1, 0.1, -0.1, 0.1, 0.1, 0.1, -0.002, -0.002, -0.005)^\top$ and $\sigma^* = 0.1$. We set $M = 5 \times N$ and used the design matrices given by

$$Z = Z_r \quad \text{and} \quad \tilde{Z} = \begin{pmatrix} Z_r \\ Z_r \\ \cdots \\ Z_r \end{pmatrix} + \lambda \begin{pmatrix} I_9 \\ O_{(M-9),9} \end{pmatrix},$$

where Z_r was given randomly. We changed the value of λ to each element in $\{50, 100, 150, 200\}$.

We considered the full model $\{\phi_N(x^{(N)}; Z\theta, \sigma^2 I_N) \times \phi_M(y^{(M)}; \tilde{Z}\theta, \sigma^2 I_M) :$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Table 4.4: The means (with the standard deviations) of negative log predictive densities in the settings with $N = 50$ and $M = 250$. The lowest value in each row is underlined.

λ	AIC	PIC	PIC_{BS}	MSPIC	MSPIC_{BS}
50	−186 (12.5)	−184 (12.3)	−190 (12.2)	544 (3.25)	<u>−194</u> (11.2)
100	−142 (22.9)	−107 (136)	−162 (27.9)	750 (3.37)	<u>−188</u> (10.2)
150	−91.1 (29.0)	−82.1 (17.1)	−124 (46.5)	871 (2.76)	<u>−182</u> (18.1)
200	−50.0 (47.8)	−18.4 (115)	−91.5 (64.6)	957 (3.42)	<u>−178</u> (23.6)

Table 4.5: The means (with the standard deviations) of negative log predictive densities in the settings with $N = 100$ and $M = 500$. The lowest value in each row is underlined.

λ	AIC	PIC	PIC_{BS}	MSPIC	MSPIC_{BS}
50	−408 (15.3)	−406 (14.8)	−412 (16.2)	−303 (353)	<u>−419</u> (15.9)
100	−360 (29.1)	−348 (19.7)	−375 (28.8)	1220 (323)	<u>−405</u> (22.2)
150	−302 (45.2)	−276 (13.8)	−337 (54.4)	1530 (5.63)	<u>−401</u> (27.4)
200	−238 (64.7)	−198 (27.8)	−296 (84.3)	1700 (5.67)	<u>−399</u> (31.9)

$\theta \in \mathbb{R}^9, \sigma > 0\}$ and 511 sub-models obtained by setting some components of θ equal to zero. By denoting the design matrix in the m-th model by Z_m, the m-th sub-model is $\{\phi_N(x^{(N)}; Z_m\theta_m, \sigma^2 I_N) \times \phi_M(y^{(M)}; \tilde{Z}_m\theta_m, \sigma^2 I_M) : \theta_m \in \mathbb{R}^{d_m}, \sigma > 0\}$. We call the full model the 512-th model for convenience. We denote Z by Z_{512}, $\tilde{Z}$ by $\tilde{Z}_{512}$, and the dimension of the full model by $d_{512} + 1 (= 10)$.

We examined a setting with $N = 50$ and $M = 250$ and a setting with $N = 100$ and $M = 500$. We compared the values of negative log Bayesian predictive densities based on $\pi(\theta_m, \sigma) = 1/\sigma$ using the AIC-best,

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

the PIC-best, the $\mathrm{PIC}_{\mathrm{BS}}$-best, the MSPIC-best, and the $\mathrm{MSPIC}_{\mathrm{BS}}$-best models. We used the above Bayesian distribution because it is mini-max under the Kullback–Leibler risk (see Liang and Barron (2004)). However, the choice of prior densities is asymptotically irrelevant according to Theorem 1.

Tables 4.4 and 4.5 show that $\mathrm{MSPIC}_{\mathrm{BS}}$ has the lowest value of the negative log predictive distribution. The dominance of $\mathrm{MSPIC}_{\mathrm{BS}}$ is enlarged depending on the value of λ. In contrast, MSPIC has the worst performance because of the variance and the skewness of MSPIC. These results suggest that we use the bootstrap adjustment $\mathrm{MSPIC}_{\mathrm{BS}}$ instead of MSPIC itself.

5. Discussion and Conclusion

In this paper, we have considered prediction when the distributions of current and future observations may differ with an identical unknown parameter. We have shown that the Bayesian predictive distribution has a smaller Kullback–Leibler risk than the plug-in predictive distribution when both N and M simultaneously grow to infinity. The asymptotic form of the Kullback–Leibler risk is different from that when N grows to infinity but M is 1. Based on the results, we have proposed a model selection criterion MSPIC for settings in which the true distributions of current and future observations may differ. The proposed model selection criterion MSPIC is an asymptotically unbiased estimator of the Kullback–Leibler risk of the Bayesian predictive distribution. We have also proposed a bootstrap adjustment $\mathrm{MSPIC}_{\mathrm{BS}}$. Numerical experiments show that our proposed model selection criterion $\mathrm{MSPIC}_{\mathrm{BS}}$ is effective.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Acknowledgments. The authors thank an associate editor and two reviewers for their careful reading and constructive suggestions on the manuscript. This work is supported by JSPS KAKENHI Grand Number 26280005.

Appendix

We provide the proofs of Theorems 1 and 3. The proofs consist of three parts: the connection formula (Lemma A1), the expansions of maximum likelihood estimators (Lemma A2), and the expansions of Kullback–Leibler risk $R_{\rm s}(\theta^*, q_{\pi_{\rm s}})$.

We use tensorial notations. To avoid the collision of indices, we use indices i, j, k for observation x_i, indices u, v, w for parameter θ, and indices a, b, c for parameter $\theta_{\rm s}$. We use indices κ, λ, μ for the parameter $\gamma_{\rm s}$. We denote parameter $(\theta_{\rm s}, \gamma_{\rm s})$ by $\xi_{\rm s}$ and we use indices α, β, γ for parameter $\xi_{\rm s}$. We denote the true parameter value with respect to $\xi_{\rm s}$ by $\xi_{\rm s}^*$.

We adopt the Einstein summation convention: if the same indices appear in any one term, it implies summation over that index.

We denote the (i, j)-component of $g^{(N)}(\theta)$ by $g^{(N)}_{ij}(\theta)$ and that of $\tilde{g}^{(M)}(\theta)$ by $\tilde{g}^{(M)}_{ij}(\theta)$. We denote the (α, β)-components of the Fisher information matrices $g^{(N)}(\xi_{\rm s})$ and $\tilde{g}^{(M)}(\xi_{\rm s})$ with respect to parameter $\xi_{\rm s}$ by $g^{(N)}_{\alpha\beta}(\xi_{\rm s})$ and $\tilde{g}^{(M)}_{\alpha\beta}(\xi_{\rm s})$, respectively. We denote the (a, b)-component of the $d_{\rm s} \times d_{\rm s}$-dimensional top left sub-matrix with respect to $\theta_{\rm s}$ of the Fisher information matrix $g^{(N)}_{\alpha\beta}(\xi_{\rm s})$ by $g^{(N)}_{{\rm s},ab}(\theta_{\rm s})$ and that of $\tilde{g}^{(M)}_{\alpha\beta}(\xi_{\rm s})$ by $\tilde{g}^{(M)}_{{\rm s},ab}(\theta_{\rm s})$.

We denote the (i, j)-components of inverse Fisher information matrices ${g^{(N)}}^{-1}(\theta)$ and ${\tilde{g}^{(M)}}^{-1}(\theta)$ by $g^{(N)ij}(\theta)$ and $\tilde{g}^{(M)ij}(\theta)$, respectively. We denote

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

the (a,b)-components of inverse Fisher information matrices ${g_{\rm s}^{(N)}}^{-1}(\theta_{\rm s})$ and ${\tilde{g}_{\rm s}^{(M)}}^{-1}(\theta_{\rm s})$ by $g_{\rm s}^{(N)ab}(\theta_{\rm s})$ and $\tilde{g}_{\rm s}^{(M)ab}(\theta_{\rm s})$, respectively. Note that the $d_{\rm s} \times d_{\rm s}$ top left sub-matrix of the inverse Fisher information matrix of $g^{(N)}(\xi_{\rm s}(\theta_s,0))$ is in general not identical to ${g_{\rm s}^{(N)}}^{-1}(\theta_{\rm s})$.

The joint distribution of $(x^{(N)},y^{(M)})$ is denoted by $r_{\theta^*}(x^{(N)},y^{(M)})$. In our setting, distribution $r_{\theta^*}(x^{(N)},y^{(M)})$ is equal to $p_{\theta^*}(x^{(N)}) \times q_{\theta^*}(y^{(M)})$. We denote Fisher information matrices of $r_\theta(x^{(N)},y^{(M)})$ and $r_{\theta_{\rm s}}(x^{(N)},y^{(M)})$ by $\bar{g}^{(N+M)}(\theta)$ and $\bar{g}_{\rm s}^{(N+M)}(\theta_{\rm s})$, respectively. Note that $\bar{g}^{(N+M)}(\theta) = g^{(N)}(\theta) + \tilde{g}^{(M)}(\theta)$. We denote $g_{u\alpha}^{(N)}\frac{\partial \xi_{\rm s}^\alpha}{\partial \theta^u}$ by $g_{au}^{(N)}$ and use $\bar{g}_{at}^{(N+M)}$ and $\tilde{g}_{at}^{(M)}$ in the same manner. We denote the maximum likelihood estimator of $r_\theta(x^{(N)},y^{(M)})$ by $\hat{\theta}(x^{(N)},y^{(M)})$ and the maximum likelihood estimator of $r_{\theta(\theta_{\rm s},0)}(x^{(N)},y^{(M)})$ by $\hat{\theta}_{\rm s}(x^{(N)},y^{(M)})$. We denote the (a,b)-components of the observed Fisher information matrices of $p_{\theta_{\rm s}}(x^{(N)})$ and $r_{\theta_{\rm s}}(x^{(N)},y^{(M)})$ by $\hat{g}_{{\rm s},ab}^{(N)}(\hat{\theta}_{\rm s}(x^{(N)}))$ and $\hat{g}_{{\rm s},ab}^{(N+M)}(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}))$, respectively.

We denote the m-projections of $p_{\theta^*}(x^{(N)})$, $q_{\theta^*}(y^{(M)})$, and $r_{\theta^*}(x^{(N)},y^{(M)})$ into $\{p_{\theta_{\rm s}}(x^{(N)}) : \theta_{\rm s} \in \Theta_{\rm s}\}$, $\{q_{\theta_{\rm s}}(y^{(M)}) : \theta_{\rm s} \in \Theta_{\rm s}\}$, and $\{r_{\theta_{\rm s}}(x^{(N)},y^{(M)}) : \theta_{\rm s} \in \Theta_{\rm s}\}$ by $\theta_{\rm s}^{(p)}$, $\theta_{\rm s}^{(q)}$, and $\theta_{\rm s}^{(r)}$, respectively. For example, $\theta_{\rm s}^{(p)}$ is defined by

$$\theta_{\rm s}^{(p)} = \underset{\theta_{\rm s}\in\Theta_{\rm s}}{\rm argmax} \int p_{\theta^*}(x^{(N)}) \log \frac{p_{\theta^*}(x^{(N)})}{p_{\theta_{\rm s}}(x^{(N)})} {\rm d}x^{(N)}.$$

See Chapter 3 of Amari (1985) for details such as the existence of the m-projection.

In this appendix, $\theta^{(p)}$ and $\theta^{(r)}$ are used instead of $\theta(\theta_{\rm s}^{(p)},0)$ and $\theta(\theta_{\rm s}^{(r)},0)$, respectively. $\xi_{\rm s}^{(p)}$ is used instead of $\xi_{\rm s}(\theta_{\rm s}^{(p)},0)$. θ_0 and $\xi_{{\rm s},0}$ are used instead of $\theta(\theta_{{\rm s},0},0)$ and $\xi_{\rm s}(\theta_{\rm s},0)$, respectively.

We denote the stochastic large and small orders with respect to the

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

distribution with parameter θ by O_θ and o_θ, respectively. We denote the expectation of $x^{(N)}$ and $y^{(M)}$ with respect to $r_\theta(x^{(N)}, y^{(M)})$ by E_θ.

First, we derive equalities with respect to m-projections $\theta_{\mathrm{s}}^{(p)}$ and $\theta_{\mathrm{s}}^{(r)}$. Note that under parameterization θ assumption (2.1) is given by

$$\theta^{*u} - \theta^u(\theta_{\mathrm{s},0}, 0) = \frac{\partial \theta^u}{\partial \xi_{\mathrm{s}}^\alpha}(\xi_{\mathrm{s}}^*)\frac{h_{\mathrm{s}}^\alpha}{\sqrt{N}} + \mathrm{o}(1/\sqrt{N}) \text{ for } u = 1, \ldots, d. \tag{1}$$

We denote $\frac{\partial \theta^u}{\partial \xi_{\mathrm{s}}^\alpha}(\xi_{\mathrm{s}}^*)h_{\mathrm{s}}^\alpha$ by h_{s}^u.

Lemma A1. *For $a = 1, \ldots, d_{\mathrm{s}}$, we have*

$$\frac{h_{\mathrm{s}}^a}{\sqrt{N}} = -g_{\mathrm{s}}^{(N)ab}(\theta_{\mathrm{s},0})g_{b\kappa}^{(N)}(\xi_{\mathrm{s},0})\frac{h_{\mathrm{s}}^\kappa}{\sqrt{N}} + \mathrm{O}(1/N). \tag{2}$$

For $a = 1, \ldots, d_{\mathrm{s}}$, we have

$$\theta_{\mathrm{s}}^{(p)a} - \theta_{\mathrm{s},0}^a = 0 + \mathrm{O}(1/N), \tag{3}$$

$$\theta_{\mathrm{s}}^{(r)a} - \theta_{\mathrm{s},0}^a = \bar{g}_{\mathrm{s}}^{(N+M)ab}(\theta_{\mathrm{s},0})\bar{g}_{bu}^{(N+M)}(\theta_0)\frac{h_{\mathrm{s}}^u}{\sqrt{N}} + \mathrm{O}(1/N). \tag{4}$$

Proof. We have

$$p_{\theta^*}(x^{(N)}) = p_{\theta_0}(x^{(N)})\left[1 + \partial_u \log p_{\theta_0}(x^{(N)})\frac{h^u}{\sqrt{N}} + \mathrm{O}_{\theta_0}(1/\sqrt{N})\right], \tag{5}$$

$$q_{\theta^*}(y^{(M)}) = q_{\theta_0}(y^{(M)})\left[1 + \partial_s \log q_{\theta_0}(y^{(M)})\frac{h^u}{\sqrt{N}} + \mathrm{O}_{\theta_0}(1/\sqrt{N})\right]. \tag{6}$$

Consider the equation

$$\mathrm{E}_{\theta^*}[\partial_\kappa \log p_{\theta^*}(x^{(N)})/\sqrt{N}] = 0 \text{ for } \kappa = 1, \ldots, d - d_{\mathrm{s}}. \tag{7}$$

From (5) and the Taylor expansion of $p_\theta(x^{(N)})$ around θ_0, we have

$$\mathrm{E}_{\theta^*}[\partial_\kappa \log p_{\theta^*}(x^{(N)})/\sqrt{N}] = g_{u\kappa}^{(N)}(\theta_0)\frac{\theta^{*u} - \theta_0^u}{\sqrt{N}} + \mathrm{E}_{\theta_0}[\mathrm{O}_{\theta_0}(1)].$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Thus, we obtain (2).

Consider the definition of $\theta^{(p)}$ and $\theta^{(r)}$. We have

$$\frac{1}{\sqrt{N}}\mathrm{E}_{\theta^*}\left[\partial_a \log p_{\theta^{(p)}}(x^{(N)})\right] = 0, \tag{8}$$

$$\frac{1}{\sqrt{N}}\mathrm{E}_{\theta^*}\left[\partial_a \log r_{\theta^{(r)}}(x^{(N)}, y^{(M)})\right] = 0. \tag{9}$$

From the independence of $x_1, \ldots, x_N, x_{N+1}, \ldots, x_{N+M}$, equations (5) and (6), and from the Taylor expansions of $p_\theta(x^{(N)})$ and $r_\theta(x^{(N)}, y^{(M)})$ around θ_0, we have

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathrm{E}_{\theta^*}[\partial_a \log p_{\theta^{(p)}}(x^{(N)})] =& g^{(N)}_{au}(\theta_0)\frac{h^u_{\mathrm{s}}}{N} - g^{(N)}_{ab}(\theta_0)\frac{\theta^{(p)b}_{\mathrm{s}} - \theta^b_{\mathrm{s},0}}{\sqrt{N}} \\ &+ \mathrm{E}_{\theta_0}[\mathrm{O}_{\theta_0}(||\theta^{(p)} - \theta_0||^2)] + \mathrm{E}_{\theta_0}[\mathrm{O}_{\theta_0}(1/N)],\end{aligned} \tag{10}$$

$$\begin{aligned}\frac{1}{\sqrt{N}}\mathrm{E}_{\theta^*}[\partial_a \log r_{\theta^{(r)}}(x^{(N)}, y^{(M)})] = \bar{g}^{(N+M)}_{au}(\theta_0)\frac{h^u_{\mathrm{s}}}{N} - \bar{g}^{(N+M)}_{ab}(\theta_0)\frac{\theta^{(r)b}_{\mathrm{s}} - \theta^b_{\mathrm{s},0}}{\sqrt{N}} \\ + \mathrm{E}_{\theta_0}[\mathrm{O}_{\theta_0}(||\theta^{(r)} - \theta_0||^2)] + \mathrm{E}_{\theta_0}[\mathrm{O}_{\theta_0}(1/N)].\end{aligned} \tag{11}$$

By substituting (11) into (9), we obtain (4). By substituting (10) and (2) into (8), we obtain (3).

□

Next, we derive the asymptotic linear forms of the maximum likelihood estimators.

Lemma A2. *For $a = 1, \ldots, d_{\mathrm{s}}$, we have*

$$\hat{\theta}^a_{\mathrm{s}}(x^{(N)}, y^{(M)}) - \theta^{(r)a}_{\mathrm{s}} = \bar{g}^{(N+M)ab}_{\mathrm{s}}(\theta^{(r)}_{\mathrm{s}})\partial_b \log r_{\theta^{(r)}_{\mathrm{s}}}(x^{(N)}, y^{(M)}) + \mathrm{O}_{\theta^*}(1/N), \tag{12}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

$$\hat{\theta}^a_{\mathrm{s}}(x^{(N)}) - \theta^{(p)a}_{\mathrm{s}} = g^{(N)ab}_{\mathrm{s}}(\theta^{(p)}_{\mathrm{s}})\partial_b \log p_{\theta^{(p)}_{\mathrm{s}}}(x^{(N)}) + \mathrm{O}_{\theta^*}(1/N). \tag{13}$$

For $u = 1, \ldots, d$, we have

$$\hat{\theta}^u(x^{(N)}, y^{(M)}) - \theta^{*u} = \bar{g}^{(N+M)uv}(\theta^*)\partial_v \log r_{\theta^*}(x^{(N)}, y^{(M)}) + \mathrm{O}_{\theta^*}(1/N), \tag{14}$$

$$\hat{\theta}^u(x^{(N)}) - \theta^{*u} = g^{(N)uv}(\theta^*)\partial_v \log p_{\theta^*}(x^{(N)}) + \mathrm{O}_{\theta^*}(1/N). \tag{15}$$

Proof. Consider the estimative equation

$$\partial_a \log r_{\theta_{\mathrm{s}}=\hat{\theta}_{\mathrm{s}}(x^{(N)}, y^{(M)})}(x^{(N)}, y^{(M)}) = 0. \tag{16}$$

From the Taylor expansion around $\theta^{(r)}_{\mathrm{s}}$ and the Central Limit Theorem for $\partial_{ab} \log r_{\theta_{\mathrm{s}}}(x^{(N)}, y^{(M)})$, we have

$$\begin{aligned}
&\partial_a \log r_{\theta_{\mathrm{s}}=\hat{\theta}_{\mathrm{s}}(x^{(N)}, y^{(M)})}(x^{(N)}, y^{(M)}) \\
&\quad = \partial_a \log r_{\theta^{(r)}_{\mathrm{s}}}(x^{(N)}, y^{(M)}) - \bar{g}^{(N+M)}_{\mathrm{s},ab}(\theta^{(r)}_{\mathrm{s}})\{\hat{\theta}^b_{\mathrm{s}}(x^{(N)}, y^{(M)}) - \theta^{(r)b}_{\mathrm{s}}\} + \mathrm{O}_{\theta_0}(1).
\end{aligned}$$

Then, we have (12). Equation (14) immediately follows from the estimative equation of $\hat{\theta}$. For example, see Theorem 5.39 in van der Vaart (1998). Likewise, we have (13) and (15). □

Proof of Theorem 1. First, from (1), (4), and (12), we have

$$\begin{aligned}
&\theta^u(\hat{\theta}_{\mathrm{s}}(x^{(N)}, y^{(M)}), 0) - \theta^{*u} \\
&\quad = \frac{\partial \theta^u}{\partial \theta^a_{\mathrm{s}}}(\theta^{(r)})\bar{g}^{(N+M)}_{\mathrm{s}}(\theta^{(r)}_{\mathrm{s}})\partial_b \log r_{\theta^{(r)}_{\mathrm{s}}}(x^{(N)}, y^{(M)}) \\
&\qquad + \frac{\partial \theta^u}{\partial \theta^a_{\mathrm{s}}}(\theta_0)\bar{g}^{(N+M)ab}_{\mathrm{s}}(\theta_{\mathrm{s},0})\bar{g}^{(N+M)}_{bv}(\theta_0)\frac{h^v_{\mathrm{s}}}{\sqrt{N}} - \frac{h^u_{\mathrm{s}}}{\sqrt{N}} + \mathrm{O}_{\theta^*}(1/N),
\end{aligned} \tag{17}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

$$\theta^u(\hat{\theta}_{\rm s}(x^{(N)}),0)-\theta^{*u} = \frac{\partial\theta^u}{\partial\theta^a_{\rm s}}(\theta^{(p)})g^{(N)ab}_{\rm s}(\theta^{(p)}_{\rm s})\partial_b \log p_{\theta^{(p)}_{\rm s}}(x^{(N)}) - \frac{h^u_{\rm s}}{\sqrt{N}} + {\rm O}_{\theta^*}(1/N). \tag{18}$$

Second, consider the decomposition of the Kullback–Leibler risk,

$$R_{\rm s}(\theta^*, q_{\pi_{\rm s}}) = {\rm E}_{\theta^*}\left[\log\frac{r_{\theta^*}(x^{(N)},y^{(M)})}{r_{\pi_{\rm s}}(x^{(N)},y^{(M)})}\right] - {\rm E}_{\theta^*}\left[\log\frac{p_{\theta^*}(x^{(N)})}{p_{\pi_{\rm s}}(x^{(N)})}\right], \tag{19}$$

where $r_{\pi_{\rm s}}(x^{(N)},y^{(M)})$ is the marginal distribution of $x^{(N)}$ and $y^{(M)}$ and $p_{\pi_{\rm s}}(x^{(N)})$ is the marginal distribution of $x^{(N)}$. Using the marginal expansions of $r_{\pi_{\rm s}}(x^{(N)},y^{(M)})$ and $p_{\pi_{\rm s}}(x^{(N)})$ given at p.117 in Ghosh et al. (2006), decomposition (19) is given by

$$\begin{aligned} R_{\rm s}(\theta^*, q_{\pi_{\rm s}}) =& {\rm E}_{\theta^*}\left[\log\frac{r_{\theta^*}(x^{(N)},y^{(M)})}{r_{\hat{\theta}_{\rm s}(x^{(N)},y^{(M)})}(x^{(N)},y^{(M)})}\right] - {\rm E}_{\theta^*}\left[\log\frac{p_{\theta^*}(x^{(N)})}{p_{\hat{\theta}_{\rm s}(x^{(N)})}(x^{(N)})}\right] \\ &+ {\rm E}_{\theta^*}\left[\frac{1}{2}\log\frac{|\hat{g}^{(N+M)}_{\rm s}(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}))|}{|\hat{g}^{(N)}_{\rm s}(\hat{\theta}_{\rm s}(x^{(N)}))|}\right] \\ &- {\rm E}_{\theta^*}\left[\log\frac{\pi_{\rm s}(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}))}{\pi_{\rm s}(\hat{\theta}_{\rm s}(x^{(N)}))}\right] + {\rm o}(1). \end{aligned} \tag{20}$$

Consider the first term in (20). Using the Taylor expansion, we expand the negative of the first term as

$$\begin{aligned} &{\rm E}_{\theta^*}\left[\log\frac{r_{\hat{\theta}_{\rm s}(x^{(N)},y^{(M)})}(x^{(N)},y^{(M)})}{r_{\theta^*}(x^{(N)},y^{(M)})}\right] \\ &= {\rm E}_{\theta^*}\left[\partial_u \log r_{\theta^*}(x^{(N)},y^{(M)})\{\theta^u(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}),0)-\theta^{*u}\}\right] \\ &\quad + \frac{1}{2}{\rm E}_{\theta^*}\left[\partial_{uv}\log r_{\theta^*}(x^{(N)},y^{(M)})\right. \\ &\quad \left.\times\{\theta^u(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}),0)-\theta^{*u}\}\{\theta^v(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}),0)-\theta^{*v}\}\right] + {\rm o}(1). \end{aligned} \tag{21}$$

From (17) and from Lemma A1, we expand the first term in (21) as

$${\rm E}_{\theta^*}\left[\partial_u \log r_{\theta^*}(x^{(N)},y^{(M)})\left\{\theta^u(\hat{\theta}_{\rm s}(x^{(N)},y^{(M)}),0)-\theta^{*u}\right\}\right] = d_{\rm s} + {\rm o}(1), \tag{22}$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

and we expand the second term in (21) as

$$\begin{aligned}
&\frac{1}{2}\mathrm{E}_{\theta^*}[\partial_{uv}\log r_{\theta^*}(x^{(N)},y^{(M)})\{\hat{\theta}^u_{\mathrm{s}}(x^{(N)},y^{(M)})-\theta^{*u}\}\{\hat{\theta}^v_{\mathrm{s}}(x^{(N)},y^{(M)})-\theta^{*v}\}]\\
&\quad=-\frac{1}{2}\bar{g}^{(N+M)}_{uv}(\theta^{(p)})\frac{h^u_{\mathrm{s}}h^v_{\mathrm{s}}}{N}+\frac{1}{2}\bar{g}^{(N+M)ab}_{\mathrm{s}}(\theta^{(p)}_{\mathrm{s}})\bar{g}^{(N+M)}_{au'}(\theta^{(p)})\bar{g}^{(N+M)}_{bv'}(\theta^{(p)})\frac{h^{u'}_{\mathrm{s}}h^{v'}_{\mathrm{s}}}{N}\\
&\qquad-\frac{1}{2}d_{\mathrm{s}}+\mathrm{o}(1). \qquad (23)
\end{aligned}$$

By combining (22) and (23), we obtain

$$\begin{aligned}
&\mathrm{E}_{\theta^*}\left[\log\frac{r_{\hat{\theta}_{\mathrm{s}}(x^{(N)},y^{(M)})}(x^{(N)},y^{(M)})}{r_{\theta^*}(x^{(N)},y^{(M)})}\right]\\
&\quad=-\frac{1}{2}\left[\bar{g}^{(N+M)}_{uv}(\theta^{(p)})-\bar{g}^{(N+M)ab}_{\mathrm{s}}(\theta^{(p)}_{\mathrm{s}})\bar{g}^{(N+M)}_{au}(\theta^{(p)})\bar{g}^{(N+M)}_{bv}(\theta^{(p)})\right]\frac{h^u_{\mathrm{s}}h^v_{\mathrm{s}}}{N}\\
&\qquad+\frac{1}{2}d_{\mathrm{s}}+\mathrm{o}(1). \qquad (24)
\end{aligned}$$

Consider the second term in (20). From the Taylor expansion around θ^*, we expand the negative of the second term in (20) as

$$\begin{aligned}
&\mathrm{E}_{\theta^*}\left[\log\frac{p_{\hat{\theta}_{\mathrm{s}}(x^{(N)})}(x^{(N)})}{p_{\theta^*}(x^{(N)})}\right]\\
&\quad=\mathrm{E}_{\theta^*}[\partial_u\log p_{\theta^*}(x^{(N)})\{\theta^u(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*u}\}]+\frac{1}{2}\mathrm{E}_{\theta^*}[\partial_{uv}\log p_{\theta^*}(x^{(N)})\\
&\qquad\times\{\theta^u(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*u}\}\{\theta^v(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*v}\}]+\mathrm{o}(1). \qquad (25)
\end{aligned}$$

From (18), we have

$$\mathrm{E}_{\theta^*}\left[\partial_u\log p_{\theta^*}(x^{(N)})\{\theta^u(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*u}\}\right]=d_{\mathrm{s}}+\mathrm{o}(1), \qquad (26)$$

$$\begin{aligned}
&\mathrm{E}_{\theta^*}[g^{(N)}_{uv}(\theta^*)\{\theta^u(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*u}\}\{\theta^v(\hat{\theta}_{\mathrm{s}}(x^{(N)}),0)-\theta^{*v}\}]\\
&\qquad=g^{(N)}_{uv}(\theta^*)\frac{h^u_{\mathrm{s}}h^v_{\mathrm{s}}}{N}+d_{\mathrm{s}}+\mathrm{o}(1). \qquad (27)
\end{aligned}$$

By substituting (26) and (27) into (25), we have

$$\mathrm{E}_{\theta^*}\left[\log\frac{p_{\hat{\theta}_{\mathrm{s}}(x^{(N)})}(x^{(N)})}{p_{\theta^*}(x^{(N)})}\right]=-\frac{1}{2}g^{(N)}_{uv}(\theta^*)\frac{h^u_{\mathrm{s}}h^v_{\mathrm{s}}}{N}+\frac{1}{2}d_{\mathrm{s}}+\mathrm{o}(1). \qquad (28)$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

The Taylor expansions around $\theta^{(p)}$ and Lemma A1 show that the third and fourth terms in (20) are equal to $(1/2)\log(|\bar{g}_{\mathrm{s}}^{(N+M)}(\theta_{\mathrm{s},0})|/|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})|) + \mathrm{o}(1)$.

Thus, from (24) and (28), the Kullback–Leibler risk $R_{\mathrm{s}}(\theta^*, q_{\pi_{\mathrm{s}}})$ is expanded as

$$\begin{aligned} &R_{\mathrm{s}}(\theta^*, q_{\pi_{\mathrm{s}}}) \\ &\quad = \frac{1}{2N}\left[\bar{g}_{uv}^{(N+M)}(\theta^*) - g_{uv}^{(N)}(\theta^*) - \bar{g}_{\mathrm{s}}^{(N+M)ab}(\theta_{\mathrm{s}}^{(p)})\bar{g}_{ua}^{(N+M)}(\theta^{(p)})\bar{g}_{vb}^{(N+M)}(\theta^{(p)})\right] \\ &\qquad \times h_{\mathrm{s}}^u h_{\mathrm{s}}^v + \frac{1}{2}\log\frac{|\bar{g}_{\mathrm{s}}^{(N+M)}(\theta_{\mathrm{s},0})|}{|g_{\mathrm{s}}^{(N)}(\theta_{\mathrm{s},0})|} + \mathrm{o}(1). \end{aligned} \tag{29}$$

Note that this is invariant up to $\mathrm{o}(1)$ under reparameterization of θ.

Finally, to complete the proof of Theorem 1, we show that

$$P_{\alpha\beta}h_{\mathrm{s}}^{\alpha}h_{\mathrm{s}}^{\beta}/N = G_{\mathrm{s},\alpha\beta}^{(N)}h_{\mathrm{s}}^{\alpha}h_{\mathrm{s}}^{\beta}/N + \mathrm{o}(1), \tag{30}$$

where P is a matrix whose (α,β)-component is given by

$$P_{\alpha\beta} := \bar{g}_{\alpha\beta}^{(N+M)}(\xi_{\mathrm{s}}^*) - g_{\alpha\beta}^{(N)}(\xi_{\mathrm{s}}^*) - \bar{g}_{\mathrm{s}}^{(N+M)ab}(\theta_{\mathrm{s}}^{(p)})\bar{g}_{a\alpha}^{(N+M)}(\xi_{\mathrm{s}}^{(p)})\bar{g}_{b\beta}^{(N+M)}(\xi_{\mathrm{s}}^{(p)}).$$

From Lemma A1 and from the evaluations of $P_{ab}h_{\mathrm{s}}^a h_{\mathrm{s}}^b$, $P_{\kappa\lambda}h_{\mathrm{s}}^{\kappa}h_{\mathrm{s}}^{\lambda}$, and $P_{a\kappa}h_{\mathrm{s}}^a h_{\mathrm{s}}^{\kappa}$, we obtain

$$\begin{aligned} P_{\alpha\beta}h_{\mathrm{s}}^{\alpha}h_{\mathrm{s}}^{\beta} = \{&\tilde{g}_{\kappa\lambda}^{(M)}(\xi_{\mathrm{s}}^*) + g_{\mathrm{s}}^{(N)ab}(\theta_{\mathrm{s},0})g_{a\kappa}^{(N)}(\xi_{\mathrm{s}}^*)g_{b\lambda}^{(N)}(\xi_{\mathrm{s}}^*) \\ &- \bar{g}_{\mathrm{s}}^{(N+M)ab}(\theta_{\mathrm{s},0})\bar{g}_{a\kappa}^{(N+M)}(\xi_{\mathrm{s}}^*)g_{b\lambda}^{(N+M)}(\xi_{\mathrm{s}}^*)\}h_{\mathrm{s}}^{\kappa}h_{\mathrm{s}}^{\lambda} + \mathrm{o}(N). \end{aligned} \tag{31}$$

By applying the Sherman–Morisson–Woodbury identity to matrix $G_{\mathrm{s}}^{(N)}$, we have

$$G_{\mathrm{s}}^{(N)} = \tilde{g}^{(M)}(\xi_{\mathrm{s}}^*) - \tilde{g}^{(M)}(\xi_{\mathrm{s}}^*)\begin{pmatrix} \bar{g}_{\mathrm{s}}^{(N+M)^{-1}}(\theta_{\mathrm{s},0}) & O_{d_{\mathrm{s}},(d-d_{\mathrm{s}})} \\ O_{(d-d_{\mathrm{s}}),d_{\mathrm{s}}} & O_{(d-d_{\mathrm{s}}),(d-d_{\mathrm{s}})} \end{pmatrix}\tilde{g}^{(M)}(\xi_{\mathrm{s}}^*).$$

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Through the evaluations of $G^{(N)}_{\mathrm{s},ab}h\ a_{\mathrm{s}}h^b_{\mathrm{s}}$, $G^{(N)}_{\mathrm{s},a\kappa}h^a_{\mathrm{s}}h^\kappa_{\mathrm{s}}$, and $G^{(N)}_{\mathrm{s},\kappa\lambda}h^\kappa_{\mathrm{s}}h^\lambda_{\mathrm{s}}$, we obtain

$$G^{(N)}_{\mathrm{s},\alpha\beta}h^\alpha_{\mathrm{s}}h^\beta_{\mathrm{s}} = \Big\{\tilde{g}^{(M)}_{\kappa\lambda}(\xi^*_{\mathrm{s}}) + g^{(N)}_{a\kappa}(\xi^*_{\mathrm{s}})g^{(N)}_{\mathrm{s}}(\theta_{\mathrm{s},0})g^{(N)}_{b\lambda}(\xi^*_{\mathrm{s}}) - \bar{g}^{(N+M)}_{a\kappa}(\xi^*_{\mathrm{s}})g^{(N+M)}_{\mathrm{s}}(\theta_{\mathrm{s},0})\bar{g}^{(N+M)}_{b\lambda}(\xi^*_{\mathrm{s}})\Big\}h^\kappa_{\mathrm{s}}h^\lambda_{\mathrm{s}} + \mathrm{o}(N).$$

Thus, we obtain (30). □

Proof of Theorem 3. Since $\hat{h}^\alpha_{\mathrm{s}}$ is decomposed as

$$\frac{\hat{h}^\alpha_{\mathrm{s}}}{\sqrt{N}} = g^{(N)\alpha\beta}(\xi^*_{\mathrm{s}})\partial_\beta \log p_{\xi^*_{\mathrm{s}}}(x^{(N)}) + \frac{h^\alpha_{\mathrm{s}}}{\sqrt{N}} - \delta^\alpha_a g^{(N)ab}_{\mathrm{s}}(\theta_{\mathrm{s},0})\partial_b \log p_{\theta_{\mathrm{s},0}}(x^{(N)}) + \mathrm{O}(1/N), \tag{32}$$

where δ^α_a is 1 if $\alpha = a$ and otherwise 0, the expectation of $\hat{G}^{(N)}_{\mathrm{s},\alpha\beta}\hat{h}^\alpha_{\mathrm{s}}\hat{h}^\beta_{\mathrm{s}}/N$ is given by

$$\mathrm{E}_{\theta^*}[\hat{G}^{(N)}_{\mathrm{s},\alpha\beta}\hat{h}^\alpha\hat{h}^\beta]/N = G^{(N)}_{\mathrm{s},\alpha\beta}\frac{h^\alpha_{\mathrm{s}}h^\beta_{\mathrm{s}}}{N} + G^{(N)}_{\mathrm{s},\alpha\beta}g^{(N)\alpha\beta}(\xi^*_{\mathrm{s}}) - G^{(N)}_{\mathrm{s},ab}g^{(N)ab}_{\mathrm{s}}(\theta_{\mathrm{s},0}) + \mathrm{o}(1). \tag{33}$$

Thus, we complete the proof. □

Bibliography

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In *Proc. of the 2nd International Symposium of Information Theory*. 267–281. Budapest: Akademiai Kiado.

Akaike, H. (1980). On the use of the predictive likelihood of a Gaussian model. *Ann. Inst. Statist. Math.* **32**, 311–324.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Amari, S. (1985). *Differential-Geometrical Methods in Statistics*. Springer.

Claeskens, G. and Hjort, N. L. (2003). The focused information criterion. *J. Amer. Statist. Assoc.* **98**, 900–916.

Fushiki, T. (2005). Bootstrap prediction and bayesian prediction under misspecified models. *Bernoulli* **11**, 747–758.

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis Theory and Methods*. Springer.

Hartigan, J. (1998). The maximum likelihood prior. *Ann. Statist.* **26**, 2083–2103.

Hjort, N. L. and Claeskens, G. (2003). Frequentist model average estimators. *J. Amer. Statist. Assoc.* **98**, 879–899.

Kitagawa, G. (1997). Information criteria for the predictive evaluation of Bayesian models. *Comm. Statist. Theory Methods* **26**, 2223–2246.

Komaki, F. (1996). On asymptotic properties of predictive distributions. *Biometrika* **83**, 299–313.

Komaki, F. (2015). Asymptotic properties of Bayesian predictive densities when the distributions of data and target variables are different. *Bayesian Anal.* **10**, 31–51.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE Trans. Inf. Theory* **50**, 2708–2726.

Statistica Sinica: Preprint
doi:10.5705/ss.202015.0380

Lv, L. and Liu, J. (2014). Model selection principles in misspecified models. *J. R. Statist. Soc. B* **76**, 141–167.

Sei, T. and Komaki, F. (2007). Bayesian prediction and model selection for locally asymptotically mixed normal models. *J. Statist. Plann. and Infer.* **137**, 2523–2534.

Shimodaira, H. (1997). Assessing the error probability of the model selection test. *Ann. Inst. Statist. Math.* **49**, 395–410.

Sin, C. and White, H. (1996). Information criteria for selecting possibly misspecified parameter models. *J. Econom.* **71**, 207–225.

Takeuchi, K. (1976). Distribution of information statistics and a criterion of model fitting. *Suri-Kagaku* **153**, 12–18. In Japanese.

van der Vaart (1998). *Asymptotic Statistics*. Cambridge University Press.

Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN

E-mail: Keisuke_Yano@mist.i.u-tokyo.ac.jp

Department of Mathematical Informatics, Graduate School of Information Science and Technology, The University of Tokyo 7-3-1 Hongo, Bunkyo-ku, Tokyo 113-8656, JAPAN

E-mail: komaki@mist.i.u-tokyo.ac.jp