MATHEMATICAL ENGINEERING
TECHNICAL REPORTS

Information Criteria for Multistep Ahead Predictions

Keisuke YANO and Fumiyasu KOMAKI

METR 2015–09 March 2015

DEPARTMENT OF MATHEMATICAL INFORMATICS
GRADUATE SCHOOL OF INFORMATION SCIENCE AND TECHNOLOGY
THE UNIVERSITY OF TOKYO
BUNKYO-KU, TOKYO 113-8656, JAPAN

WWW page: http://www.keisu.t.u-tokyo.ac.jp/research/techrep/index.html

Information criteria for multistep ahead predictions

Keisuke YANO and Fumiyasu KOMAKI
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
{keisuke_yano,komaki}@mist.i.u-tokyo.ac.jp

Abstract

We propose an information criterion for multistep ahead predictions. It is also used for extrapolations. For the derivation, we consider multistep ahead predictions under local misspecification. In the prediction, we show that Bayesian predictive distributions asymptotically have smaller Kullback–Leibler risks than plug-in predictive distributions. From the results, we construct an information criterion for multistep ahead predictions by using an asymptotically unbiased estimator of the Kullback–Leibler risk of Bayesian predictive distributions. We show the effectiveness of the proposed information criterion throughout the numerical experiments.

1 Introduction

Consider multistep ahead predictions as follows: let $x^{(N)} = (x_1, \ldots, x_N)$ be data from distribution $p(x^{(N)})$ and let $y^{(M)} = (y_1, \ldots, y_M)$ be target variables from distribution $q(y^{(M)})$. We assume that sample size M is given as the constant multiplication of sample size N, i.e., we assume that $M = cN$. We predict the distribution of the target variables on the basis of the data. Here, distributions $p(x^{(N)})$ and $q(y^{(M)})$ may be different but we assume that $x_1, \ldots, x_N, y_1, \ldots, y_M$ are independent.

For the prediction, we consider m_{full} parametric models of the distributions of the data and the target variables as follows: for $m \in \{1, \ldots, m_{\text{full}}\}$, the m-th model $\mathcal{M}_m$ is given as $\{p_m(x^{(N)}|\theta_m)q_m(y^{(M)}|\theta_m) : \theta_m \in \Theta_m\}$. Here, Θ_m is a d_m-dimensional parametric space. For simplicity, we denote parameter $\theta_{m_{\text{full}}}$ by ω, distribution $p_{m_{\text{full}}}(x^{(N)}|\omega)$ by $p(x^{(N)}|\omega)$, and distribution $q_{m_{\text{full}}}(y^{(M)}|\omega)$ by $q(y^{(M)}|\omega)$. We denote parameter space $\Theta_{m_{\text{full}}}$ by Θ and dimension $d_{m_{\text{full}}}$ by d_{full}. After the model selection, we construct the predictive distribution in the selected model.

As an example, consider the curve fitting. We obtain the values of the unknown curve at points $(z_1, \ldots, z_i, \ldots, z_N)$ and predict the distribution of the values at points $(z_{N+1}, \ldots, z_{N+j}, \ldots, z_{N+M})$. We use regression models with the basis set $\{\phi_a\}_{a=1}^{d_{\text{full}}}$: for $m \in \{1, \ldots, d_{\text{full}}\}$, for $i \in \{1, \ldots, N\}$, and for $j \in \{1, \ldots, M\}$, the i-th data and the j-th target variable in the m-th model are given by

$$x_i = \sum_{a=1}^{m} \phi_a(z_i)\theta_m^a + \epsilon_i \quad \text{and} \quad y_j = \sum_{a=1}^{m} \phi_a(z_{N+j})\theta_m^a + \epsilon_{N+j},$$

respectively. Here, $\theta_m = (\theta_m^1, \ldots, \theta_m^m)$ represents an unknown vector. Two random vectors $\epsilon = (\epsilon_1, \ldots, \epsilon_N)^\top$ and $\tilde{\epsilon} = (\epsilon_{N+1}, \ldots, \epsilon_{N+M})^\top$ are independent and distributed according to Gaussian distributions with mean zero and diagonal covariance matrices $\sigma^2 I_{N\times N}$ and $\sigma^2 I_{M\times M}$, respectively.

We measure the performance of the predictive distribution $\hat{q}$ by the Kullback–Leibler risk:

$$R(p(\cdot)q(\cdot), \hat{q}) = \int p(x^{(N)}) \int q(y^{(M)}) \log \frac{q(y^{(M)})}{\hat{q}(y^{(M)}; x^{(N)})} \mathrm{d}y^{(M)} \mathrm{d}x^{(N)}.$$

In this paper, we consider the asymptotics as the sample sizes N and M simultaneously go to infinity. Note that since $M = cN$ we consider that N goes to infinity. We show that for any smooth prior π, the Bayesian predictive distribution $q_{m,\pi}(y^{(M)}|x^{(N)})$ in submodel $\mathcal{M}_m$

$$q_{m,\pi}(y^{(M)}|x^{(N)}) = \frac{\int q_m(y^{(M)}|\theta_m)p_m(x^{(N)}|\theta_m)\pi(\theta_m)\mathrm{d}\theta_m}{\int p_m(x^{(N)}|\theta_m)\pi(\theta_m)\mathrm{d}\theta_m} \quad (1)$$

asymptotically has smaller Kullback–Leibler risk than the plug-in predictive distribution $q_m(y^{(M)}|\hat{\theta}_m(x^{(N)}))$ with the maximum likelihood estimator in submodel $\mathcal{M}_m$. Further, the Kullback–Leibler risk of the Bayesian predictive distribution varies according to the Fisher information matrices of the data and the target variables; in the *i.i.d.* settings, the risk varies according to the multiplicative constant c.

From the results, we construct an information criterion for the multistep ahead prediction by using an asymptotically unbiased estimator of the Kullback–Leibler risk of the Bayesian predictive distribution. Several numerical experiments show the performance of the proposed information criterion.

This paper is organized as follows: in Section 2, we prepare the notations and state the assumptions to be used. In Section 3, we show that Bayesian predictive distributions have smaller Kullback–Leibler risks than plug-in predictive distributions in multistep ahead predictions. In Section 4, we propose information criteria for multistep ahead predictions. By considering the variance of proposed information criteria, we propose their bootstrap adjustments. In Section 5, we show two numerical experiments: the curve fitting and the normal regression model with an unknown variance. In Section 6, we present our conclusions.

2 Notations and Assumptions

We consider that the true distributions $p(x^{(N)})$ and $q(y^{(M)})$ belong to the full model $\mathcal{M}_{m_{\mathrm{full}}}$:

$$p(x^{(N)}) = p(x^{(N)}|\omega^*) \quad \text{and} \quad q(y^{(M)}) = q(y^{(M)}|\omega^*),$$

where ω^* is a certain point in Θ. We refer to this parameter point ω^* as the true parameter point.

We consider that the full model $\mathcal{M}_{m_{\mathrm{full}}}$ contains submodel $\mathcal{M}_m$. Then, we decompose the parameter ω in the full model $\mathcal{M}_{m_{\mathrm{full}}}$ into $\omega(\theta_m, \gamma_m)$. We denote the parameterization (θ_m, γ_m) by ξ. Under parameterization ξ, we denote the true parameter point by ξ^*.

To avoid the collision of indices, we use index i, j, k for observation x_i, index s, t, u for parameter ω^s, and index a, b, c for parameter θ_m^a. We use index κ, λ, μ for parameter γ_m^κ, index α, β, γ for parameter ξ^α, and index m, n, l for submodel $\mathcal{M}_m$.

For simplicity, we denote the Kullback–Leibler risk by $R(\omega^*, \hat{q})$, i.e., the function of the true parameter point ω^* and predictive distribution $\hat{q}$. We denote the expectation with respect to the distribution with the parameter point ω by E_ω.

We consider two maximum likelihood estimators. We denote the maximum likelihood estimator of $p(x^{(N)}|\omega)$ by $\hat{\omega}(x^{(N)})$ and the restricted maximum likelihood estimator of $p(x^{(N)}|\omega(\theta_m, 0))$ by $\hat{\theta}_m(x^{(N)})$.

We consider the projection of the true parameter point into Θ_m. We denote the best approximating point of ω^* with respect to $p_m(x^{(N)}|\theta_m)$ by $\theta_m^{(p)}$. In other words, $\theta_m^{(p)}$ is defined by

$$\theta_m^{(p)} = \operatorname*{argmax}_{\theta_m \in \Theta_m} \mathrm{E}_{\omega^*}[\log p(x^{(N)}|\omega(\theta_m, 0))].$$

We denote the (i, j)-component of the Fisher information matrix of $p(x^{(N)}|\omega)$ by $g_{ij}^{(p)}(\omega)$ and that of $q(y^{(M)}|\omega)$ by $g_{ij}^{(q)}(\omega)$, and we denote the (α, β)-components of those with respect

to parameter ξ by $g^{(p)}_{\alpha\beta}(\xi)$ and $g^{(q)}_{\alpha\beta}(\xi)$, respectively. We denote the (a,b)-component of the sub-matrix with respect to θ_m of Fisher information matrix $g^{(p)}_{\alpha\beta}(\xi)$ by $g^{(p)}_{ab}(\theta_m)$ and that of $g^{(q)}_{\alpha\beta}(\xi)$ by $g^{(q)}_{ab}(\theta_m)$. We denote the sub-matrices with (a,b)-components as $g^{(p)}_{ab}(\theta_m)$ and $g^{(q)}_{ab}(\theta_m)$ by $g^{(p)}(\theta_m)$ and $g^{(q)}(\theta_m)$, respectively.

We write the upper index -1 to denote the inverse of the matrix; we denote the inverses of Fisher information matrices $g^{(p)}(\omega)$, $g^{(q)}(\omega)$, $g^{(p)}(\xi)$, and $g^{(q)}(\xi)$ by $g^{(p)-1}(\omega)$, $g^{(q)-1}(\omega)$, $g^{(p)-1}(\xi)$, and $g^{(q)-1}(\xi)$, respectively. We use the upper index for the components of the inverse of the Fisher information matrix; we denote the (i,j)-components of the inverse Fisher information matrices $g^{(p)-1}$ and $g^{(q)-1}$ by $g^{(p)ij}(\omega)$ and $g^{(q)ij}(\omega)$, respectively. We denote the (a,b)-components of the inverse Fisher information matrices $g^{(p)-1}(\theta_m)$ and $g^{(q)-1}(\theta_m)$ by $g^{(p)ab}_m(\theta_m)$ and $g^{(q)ab}_m(\theta_m)$, respectively. Note that the (a,b)-component of the inverse Fisher information matrix with (α,β)-component as $g^{(p)\alpha\beta}(\xi(\theta_m,0))$ is not generally identical to $g^{(p)ab}_m(\theta_m)$. We adopt Einstein summation convention: if the same indices appear in any one term, it implies summation over that index.

For the model selection, we consider local misspecification. The local misspecification is that the true parameter point ξ^* and submodel $\mathcal{M}_m$ satisfy the following equation:

$$\sqrt{N}\{\xi^{*\alpha} - \xi^{\alpha}(\theta^{(p)}_m, 0)\} = h^{\alpha} \quad \text{for} \quad \alpha = 1, \ldots, d_{\text{full}}. \tag{2}$$

If h vanishes, the assumption means that the true distribution is included in submodel $\mathcal{M}_m$. Thus, the assumption is an extension of the assumption that the true distribution is included in submodel $\mathcal{M}_m$. The assumption is known as local alternatives in statistical test theory. See van der Vaart (1998). The local misspecification in the model selection context is argued, for example, in Shimodaira (1997), Hjort and Claeskens (2003), and Claeskens and Hjort (2003). See also Leeb and Pötscher (2005). Note that the assumption does not depend on parameterizations: if we adopt parameterization ω, the assumption (2) is denoted by

$$\sqrt{N}\{\omega^{*s} - \omega^{s}(\theta^{(p)}_m, 0)\} = \frac{\partial \omega^s}{\partial \xi^{\alpha}}(\xi^*)h^{\alpha} + \mathrm{o}(1) \quad \text{for} \quad s = 1, \ldots, d_{\text{full}}. \tag{3}$$

In this parameterization, we denote $\frac{\partial \omega^s}{\partial \xi^{\alpha}}(\xi^*)h^{\alpha}$ in (3) by h^s.

3 Multi-step ahead predictions under local misspecification

First, we expand the Kullback–Leibler risk of the Bayesian predictive distribution in multistep ahead predictions under local misspecification. Next, we show that the Kullback–Leibler risk of the Bayesian predictive distribution is asymptotically smaller than that of the plug-in predictive distribution.

Theorem 3.1. *Assume that the true parameter point ξ^* and submodel $\mathcal{M}_m$ satisfy (2). Then, for any smooth prior π, the Kullback–Leibler risk of the Bayesian predictive distribution $q_{m,\pi}$ in submodel $\mathcal{M}_m$ is asymptotically expanded as*

$$R(\omega^*, q_{m,\pi}) = \frac{1}{2N} S_{\alpha\beta}(\xi^*)h^{\alpha}h^{\beta} + \frac{1}{2}\log\frac{|g^{(p)}(\theta^{(p)}_m) + g^{(q)}(\theta^{(p)}_m)|}{|g^{(p)}(\theta^{(p)}_m)|} + \mathrm{o}(1), \tag{4}$$

where $|\cdot|$ is a determinant and $S_{\alpha\beta}(\xi^*)$ is the (α,β)-component of the matrix given by

$$S(\xi^*) = \left(g^{(q)-1}(\xi^*) + \begin{pmatrix} g^{(p)-1}_m(\theta^{(p)}_m) & 0^{\top}_{(d_{\text{full}}-d_m)\times d_m} \\ 0_{(d_{\text{full}}-d_m)\times d_m} & 0_{(d_{\text{full}}-d_m)\times(d_{\text{full}}-d_m)} \end{pmatrix} \right)^{-1}.$$

Here, $0_{(d_{\text{full}}-d_m)\times d_m}$ is the $(d_{\text{full}}-d_m)\times d_m$-dimensional zero matrix and $0_{(d_{\text{full}}-d_m)\times(d_{\text{full}}-d_m)}$ is the $(d_{\text{full}}-d_m)\times(d_{\text{full}}-d_m)$-dimensional zero matrix.

The proof is given in the appendix. The expansion is invariant up to constant order under the reparameterization ω in the full model. See (44) in the appendix.

Remark 3.2. Note that the asymptotic Kullback–Leibler risk of the Bayesian predictive distribution does not depend on priors up to constant order. This corresponds to the fact that the asymptotic Kullback–Leibler risk of the Bayesian predictive distribution in one-step ahead predictions does not depend on priors up to the N^{-1} order. If h vanishes and if the data and the target variables are identically and identically distributed, then, $R(\omega^*, q_{m,\pi})$ is given by $d\log\{(N+M)/N\}/2$ up to constant order. In one-step ahead predictions, it is known that the asymptotic Kullback–Leibler risk of the Bayesian predictive distributions is given as $d/(2N)$ up to the N^{-1} order. The Bayesian predictive distribution $q_{m,\pi}$ is decomposed as

$$q_{m,\pi}(y^{(M)}|x^{(N)}) = q_{m,\pi}(y_M|x^{(N)}, y^{(M-1)})q_{m,\pi}(y_{M-1}|x^{(N)}, y^{(M-2)})\dots q_{m,\pi}(y_1|x^{(N)}).$$

Since the Kullback–Leibler risk of the Bayesian predictive distribution is decomposed according to the above decomposition, $R(\omega^*, q_{m,\pi})$ is also calculated as $\lim_{N\to\infty}\Sigma_{j=1}^{M} d/(2N+2j)$. This is equal to $d\log\{(N+M)/N\}/2$.

By using the above theorem, we show that the Bayesian predictive distribution has smaller Kullback–Leibler risk than the plug-in predictive distribution in the multistep ahead prediction.

Theorem 3.3. *Assume that the true parameter point ξ^* and submodel $\mathcal{M}_m$ satisfy (2). Then, for any smooth prior π, the Kullback–Leibler risk $R(\omega^*, q_{m,\pi})$ of the Bayesian predictive distribution in submodel $\mathcal{M}_m$ is smaller in constant order than the Kullback–Lebler risk $R(\omega^*, q_m(\cdot|\hat{\theta}_m))$ of the plug-in predictive distribution with the maximum likelihood estimator in submodel $\mathcal{M}_m$:*

$$\lim_{N\to\infty} R(\omega^*, q_{m,\pi}) \geq \lim_{N\to\infty} R(\omega^*, q_m(\cdot|\hat{\theta}_m)).$$

Proof. From the Taylor expansion and from (39) in the appendix, the Kullback–Leibler risk $R(\omega^*, q_m(\cdot|\hat{\theta}_m))$ is expanded as

$$\begin{aligned} R(\omega^*, q_m(\cdot|\hat{\theta}_m)) &= \frac{1}{2}g_{st}^{(q)}(\omega^*)\mathrm{E}_{\omega^*}[\{\omega^{*s} - \omega^s(\hat{\theta}_m(x^{(N)}), 0)\}\{\omega^{*t} - \omega^t(\hat{\theta}_m(x^{(N)}), 0)\}] + \mathrm{o}(1) \\ &= \frac{1}{2N}g_{\alpha\beta}^{(q)}(\xi^*)h^\alpha h^\beta + \frac{1}{2}g_m^{(q)ab}(\theta_m^{(p)})g_{ab}^{(p)}(\theta_m^{(p)}) + \mathrm{o}(1). \end{aligned}$$

Since the Fisher information matrices $g^{(p)}(\theta_m^{(p)})$ and $g^{(q)}(\theta_m^{(p)})$ are positive semidefinite, the following inequality holds:

$$\log\frac{|g^{(p)}(\theta_m^{(p)}) + g^{(q)}(\theta_m^{(p)})|}{|g^{(p)}(\theta_m^{(p)})|} \geq g_m^{(p)ab}(\theta_m^{(p)})g_{ab}^{(q)}(\theta_m^{(p)}).$$

From the inequality that $g^{(q)}(\xi^*) \succeq S$, we have

$$g_{\alpha\beta}^{(q)}(\xi^*)h^\alpha h^\beta \geq S_{\alpha\beta}h^\alpha h^\beta,$$

where the binary relation $A \succeq B$ means that $A - B$ is positive semidefinite. Thus, we complete the proof. □

Remark 3.4. This theorem implies that we should use the Bayesian predictive distribution for multistep ahead predictions instead of the plug-in predictive distribution from the viewpoint of Kullback–Leibler risk. Thus, we consider the information criteria when we use the Bayesian predictive distribution in the selected model. In one-step ahead prediction, it is well-known that the Bayesian predictive distribution has smaller Kullback–Leibler risk than the plug-in predictive distribution up to the N^{-2} order. See Komaki (1996), Hartigan (1998), and Komaki (2015). Konishi and Kitagawa (2003) construct information criteria when using the Bayesian predictive distribution in one-step ahead predictions.

Remark 3.5. The result is related to the prediction in the locally asymptotically mixed normal (LAMN) models as follows: due to the LAMN property, we consider the prediction of the target variables based on the data conditioning on the two Fisher information matrices of the data and the target variables. In our setting, we also consider the prediction of the target variables based on the data conditioning on the two Fisher information matrices of the data and the target variables. Indeed, the Kullback–Leibler risk of the Bayesian predictive distributions (4) has the same form as (2) in Sei and Komaki (2007).

4 Information criteria for multistep ahead predictions

On the basis of the results in the previous section, we construct an information criterion by using an asymptotically unbiased estimator of the Kullback–Leibler risk.

Theorem 4.1. *Let $\hat{R}(m)$ be an estimator of the Kullback–Leibler risk of the Bayesian predictive distribution in submodel $\mathcal{M}_m$ given by*

$$\begin{aligned}\hat{R}(m) &= \frac{1}{2N}\hat{S}_{\alpha\beta}\hat{h}^{\alpha}\hat{h}^{\beta} + \frac{1}{2}\hat{S}_{ab}g_m^{(p)ab}(\hat{\theta}_m) - \frac{1}{2}\hat{S}_{\alpha\beta}g^{(p)\alpha\beta}(\hat{\xi}) \\ &\quad + \frac{1}{2}\log\frac{|g^{(p)}(\hat{\theta}_m) + g^{(q)}(\hat{\theta_m})|}{|g^{(p)}(\hat{\theta}_m)|}, \end{aligned} \tag{5}$$

where $\hat{S}_{\alpha\beta}$ is the (α,β)-component of the matrix given by

$$\hat{S} = \left(g^{(q)-1}(\hat{\xi}) + \begin{pmatrix} g_m^{(p)-1}(\hat{\theta}_m) & O^{\top}_{(d_{\rm full}-d_m)\times d_m} \\ O_{(d_{\rm full}-d_m)\times d_m} & O_{(d_{\rm full}-d_m)\times(d_{\rm full}-d_m)}\end{pmatrix}\right)^{-1}$$

and for $\alpha \in \{1,\ldots,d_{\rm full}\}$, $\hat{h}^{\alpha}$ is given by $\hat{h}^{\alpha}/\sqrt{N} = \hat{\xi}^{\alpha} - \xi^{\alpha}(\hat{\theta}_m, 0)$. Assume that the true parameter point ξ^* and submodel $\mathcal{M}_m$ satisfy (2). Then, $\hat{R}(m)$ is an asymptotically unbiased estimator of the Kullback–Leibler risk $R(\omega^*, q_{m,\pi})$.

The proof is given in the appendix.

From Theorem 4.1, we propose the following model selection criterion as the multistep predictive information criterion (MSPIC):

$$\begin{aligned}\mathrm{MSPIC}(m) &= 2\hat{R}(m) \\ &= \frac{1}{N}\hat{S}_{\alpha\beta}\hat{h}^{\alpha}\hat{h}^{\beta} + \hat{S}_{ab}g_m^{(p)ab}(\hat{\theta}_m) - \hat{S}_{\alpha\beta}g^{(p)\alpha\beta}(\hat{\xi}) + \log\frac{|g^{(p)}(\hat{\theta}_m) + g^{(q)}(\hat{\theta}_m)|}{|g^{(p)}(\hat{\theta}_m)|}.\end{aligned}$$

Here, we multiply $\hat{R}(m)$ by 2 to make the definition consistent with AIC (Akaike, 1973). If two Fisher information matrices $g^{(p)}(\theta_m)$ and $g^{(q)}(\theta_m)$ are identical, MSPIC coincides with PIC (Kitagawa, 1997) when using the uniform prior and with predictive likelihood (Akaike, 1980).

We also consider the bootstrap adjustment of MSPIC. First, we generate B bootstrap samples $x_1^{(N)},\ldots,x_b^{(N)},\ldots,x_B^{(N)}$ via a parametric or non-parametric bootstrap method using the full model. Second, for each b in $\{1,\ldots,B\}$, we calculate the value of $\mathrm{MSPIC}_1(m;x_b^{(N)})$ where $\mathrm{MSPIC}_1(m;x_b^{(N)})$ is the value of

$$\frac{1}{N}\hat{S}_{\alpha\beta}\hat{h}^{\alpha}\hat{h}^{\beta} + \hat{S}_{ab}g_m^{(p)ab}(\hat{\theta}_m) - \hat{S}_{\alpha\beta}g^{(p)\alpha\beta}(\hat{\xi})$$

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

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

Consider the first three terms in the definition of MSPIC. These terms are an asymptotically unbiased estimator of $S_{\alpha\beta}h^{\alpha}h^{\beta}/N$. However, this estimator may have excessive variance because the matrix $\hat{S}$ is not equal to the asymptotic variance of $\hat{h}^{\alpha}$. To avoid the excessive variance of the estimator, we use the bootstrap method. Lv and Liu (2014) applied the bootstrap adjustment of TIC (Takeuchi, 1976).

5 Numerical experiments

We show that the proposed information criteria are effective for the multistep ahead prediction through two numerical experiments. After the model selections by AIC, PIC, MSPIC, and its bootstrap adjustment $\mathrm{MSPIC}_{\mathrm{BS}}$, we evaluate the predictive performance of the selected models as follows: the derivation of AIC is based on the plug-in predictive distribution with the maximum likelihood. In contrast, those of PIC, MSPIC, and $\mathrm{MSPIC}_{\mathrm{BS}}$ are based on the Bayesian predictive distribution. Thus, the predictive performance of the AIC-best model is evaluated by the goodness of the plug-in predictive distribution $q_m(\cdot|\hat{\theta}_m)$ in the AIC-best model. In contrast, the predictive performance of the PIC-best, the MSPIC-best, and the $\mathrm{MSPIC}_{\mathrm{BS}}$-best models is evaluated by the goodness of the Bayesian predictive distributions $q_{m,\pi}(\cdot|\cdot)$ in the PIC-best, the MSPIC-best, and the $\mathrm{MSPIC}_{\mathrm{BS}}$-best models.

We consider the empirical goodness of the predictive distribution as follows. We generate the data and the target variables R times and calculate the mean of minus log predictive densities $-\sum_{r=1}^{R}\log\hat{q}(y_r^{(M)}|x_r^{(N)})$ of each information criterion. Here, for $r=1,\ldots,R$, $x_r^{(N)}$ and $y_r^{(M)}$ are the r-th data and the r-th target variables. It is preferable that the value $-\sum_{r=1}^{R}\log\hat{q}(y_r^{(M)}|x_r^{(N)})$ is small because it is an estimator of the Kullback–Leibler risk up to the term related to the predictive distribution. We set $R=100$ in the first numerical experiment and $R=10$ in the second numerical experiment.

5.1 The extrapolation in the curve fitting

First, consider the extrapolation in the curve fitting in the introduction. For $m\in\{1,\ldots,d_{\mathrm{full}}\}$, the data and the target variables in the m-th model are given by

$$x^{(N)\top}=\Phi_m\theta_m+\epsilon_{N\times N}\quad\text{and}\quad y^{(M)\top}=\tilde{\Phi}_m\theta_m+\tilde{\epsilon}_{M\times M},$$

where Φ_m and $\tilde{\Phi}_m$ are design matrices defined by

$$\Phi_m=\begin{pmatrix}\phi_1(z_1) & \ldots & \phi_{d_m}(z_1)\\ \ldots & \ldots & \ldots\\ \phi_1(z_N) & \ldots & \phi_{d_m}(z_N)\end{pmatrix}\quad\text{and}\quad\tilde{\Phi}_m=\begin{pmatrix}\phi_1(z_{N+1}) & \ldots & \phi_{d_m}(z_{N+1})\\ \ldots & \ldots & \ldots\\ \phi_1(z_{N+M}) & \ldots & \phi_{d_m}(z_{N+M})\end{pmatrix},$$

respectively. For simplicity, we denote $\Phi_{d_{\mathrm{full}}}$, $\tilde{\Phi}_{d_{\mathrm{full}}}$, and $\theta_{d_{\mathrm{full}}}$ by Φ, $\tilde{\Phi}$, and θ, respectively. We denote the maximum likelihood estimator of θ by $\hat{\theta}$.

The information criteria AIC, PIC, and MSPIC are given by

$$\mathrm{AIC}(m) = (\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})^{\top}S_{\mathrm{AIC}}(\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})+2d_m-d_{\mathrm{full}}, \tag{6}$$

$$\mathrm{PIC}(m) = (\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})^{\top}S_{\mathrm{PIC}}(\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})+d_m\log 2+d_m-d_{\mathrm{full}}, \tag{7}$$

and

$$\begin{aligned}\mathrm{MSPIC}(m) &= (\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})^{\top}S_{\mathrm{MSPIC}}(\hat{\theta}-\begin{pmatrix}\hat{\theta}_m\\0\end{pmatrix})+\log\frac{|\Phi_m^{\top}\Phi_m+\tilde{\Phi}_m^{\top}\tilde{\Phi}_m|}{|\Phi_m^{\top}\Phi_m|}\\ &\quad+\mathrm{tr}\begin{pmatrix}\sigma^2(\Phi_m^{\top}\Phi_m)^{-1} & 0_{(d_{\mathrm{full}}-d_m)\times d_m}^{\top}\\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} & 0_{(d_{\mathrm{full}}-d_m)\times(d_{\mathrm{full}}-d_m)}\end{pmatrix}S_{\mathrm{MSPIC}}\\ &\quad-\mathrm{tr}(\Phi^{\top}\Phi)^{-1}S_{\mathrm{MSPIC}},\end{aligned} \tag{8}$$

Table 1: The mean of the minus log predictive densities when the true function is f_1 and α is 1. The lowest value in each row is underlined.

N and M	AIC	PIC	MSPIC	MSPIC_{BS}
100 and 100	−4.43	−8.71	−8.71	<u>−9.11</u>
100 and 200	−9.52	−21.84	−22.20	<u>−23.04</u>
100 and 500	−19.26	−62.33	−65.96	<u>−67.51</u>
100 and 1000	−40.93	−139.66	−150.30	<u>−152.55</u>

where S_{AIC}, S_{PIC}, and S_{MSPIC} are given by

$$S_{\text{AIC}} = \frac{1}{\sigma^2}\Phi^\top\Phi, \tag{9}$$

$$S_{\text{PIC}} = \frac{1}{\sigma^2}\left((\Phi^\top\Phi)^{-1} + \begin{pmatrix} (\Phi_m^\top\Phi_m)^{-1} & 0^\top_{(d_{\text{full}}-d_m)\times d_m} \\ 0_{(d_{\text{full}}-d_m)\times d_m} & 0_{(d_{\text{full}}-d_m)\times(d_{\text{full}}-d_m)} \end{pmatrix}\right)^{-1}, \tag{10}$$

and

$$S_{\text{MSPIC}} = \frac{1}{\sigma^2}\left((\tilde{\Phi}^\top\tilde{\Phi})^{-1} + \begin{pmatrix} (\Phi_m^\top\Phi_m)^{-1} & 0^\top_{(d_{\text{full}}-d_m)\times d_m} \\ 0_{(d_{\text{full}}-d_m)\times d_m} & 0_{(d_{\text{full}}-d_m)\times(d_{\text{full}}-d_m)} \end{pmatrix}\right)^{-1}, \tag{11}$$

respectively.

As the sets of functions $\{\phi_a\}_{a=1}^{d_{\text{full}}}$, we use trigonometric functions $\{\phi_{\text{tri},a}\}_{a=1}^{d_{\text{full}}}$:

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

For all $i \in \{1,\ldots,N+M\}$, we design z_i as $\alpha \times (i/N)$ where α is in [0,1].

We generate the data and the target variables as follows:

$$x^{(N)\top} = \begin{pmatrix} f(z_1) \\ f(z_2) \\ \cdots \\ f(z_N) \end{pmatrix} + \epsilon_{N\times N} \quad \text{and} \quad y^{(M)\top} = \begin{pmatrix} f(z_{N+1}) \\ f(z_{N+2}) \\ \cdots \\ f(z_M) \end{pmatrix} + \tilde{\epsilon}_{M\times M}.$$

In this experiment, we compare the minus log plug-in predictive distribution with the maximum likelihood estimator in the AIC-best model and the minus log Bayesian predictive distribution with the uniform prior given by

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

of the PIC-best, the MSPIC-best, and the MSPIC_{BS}-best models. Here, we denote the maximum likelihood estimator of $r_m(x^{(N)}, y^{(M)}|\theta_m)$ by $\hat{\theta}_m$.

First, we consider the setting where the true function f_1 is given by

$$\begin{aligned} f_1(z) &= 2\sin(2\pi\times z) + 0.2\sin(2\pi\times 4z) \\ &\quad + 0.1\sin(2\pi\times 8z) + 0.1\sin(2\pi\times 12z), \end{aligned}$$

where $\sigma^2 = (0.2)^2$ and $\alpha = 1.0$. We let $d_{\text{full}} = 31$. Table 1 shows that MSPIC_{BS} has the lowest value, regardless of N and M when α is 1.

Table 2: The mean of the minus log predictive densities when the true function is f_2 and α is 1. The lowest value in each row is underlined.

N and M	AIC	PIC	MSPIC	MSPIC_{BS}
100 and 100	−11.44	−13.28	−13.28	<u>−13.57</u>
100 and 200	−21.53	−28.08	−28.32	<u>−28.58</u>
100 and 500	−60.21	−79.27	−79.94	<u>−81.73</u>
100 and 1000	−116.74	−158.14	−161.33	<u>−165.81</u>

Table 3: The mean of the minus log predictive densities when the true function is f_2 and α is 0.9. The lowest value in each row is underlined.

N and M	AIC	PIC	MSPIC	MSPIC_{BS}
100 and 100	−8.91	<u>−13.48</u>	−12.98	−13.23
100 and 200	−14.88	−27.08	−26.98	<u>−27.38</u>
100 and 500	−20.99	−68.47	−70.99	<u>−72.98</u>
100 and 1000	−75.39	−154.72	−158.20	<u>−163.44</u>

Second, we consider the setting where the true function f_2 is given by

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

We set $\sigma^2 = (0.2)^2$ and $d_{\text{full}} = 16$. We consider the settings with $\alpha = 1$ and $\alpha = 0.9$. Table 2 shows that when α is 1, MSPIC_{BS} has the lowest value of the minus log predictive distribution, regardless of the ratio of N and M. Table 3 shows that when α is 0.9, MSPIC_{BS} has the lowest value except when N and M are 100 and 100, respectively.

There is difference between the first and second settings. In the first setting, the true function f_1 is included in the full model. In the second setting, the true function f_2 is not included in the full model. See Shibata (1981) for details related to the second setting. However, the experiments indicate that MSPIC_{BS} works well in both settings and that the dominance of MSPIC_{BS} is enlarged as the ratio of N and M grows.

5.2 Normal regression model with an unknown variance

Next, consider the normal regression model with an unknown variance. We consider the full model given by

$$x^{(N)\top} = \Phi\theta + \sigma\epsilon_{N\times N} \quad \text{and} \quad y^{(M)\top} = \tilde{\Phi}\theta + \sigma\tilde{\epsilon}_{M\times M},$$

respectively. Here, Φ and $\tilde{\Phi}$ are $N\times 10$ and $M\times 10$ design matrices, respectively. The parameters θ and σ are unknown. We consider 511 submodels given by the models with the restriction that some components of θ vanish. We denote the design matrix in the m-th model by Φ_m and denote the m-th model $\mathcal{M}_m$ by

$$x^{(N)\top} = \Phi_m\theta_m + \sigma\epsilon_{N\times N} \quad \text{and} \quad y^{(M)\top} = \tilde{\Phi}_m\theta_m + \sigma\tilde{\epsilon}_{M\times M},$$

respectively.

We set $N = 50$ and $M = 250$. In this setting, we generate the full design matrices given by

$$\Phi = \Phi_r \quad \text{and} \quad \tilde{\Phi} = \begin{pmatrix} \Phi_r \\ \Phi_r \\ \cdots \\ \Phi_r \end{pmatrix} + \lambda \begin{pmatrix} I_{10\times 10} \\ 0_{(M-10)\times 10} \end{pmatrix},$$

where Φ_r is given randomly and λ is the parameter. Here, $I_{10\times 10}$ is the 10×10 identity matrix and $0_{(M-10)\times 10}$ is the $(M-10) \times 10$ zero matrix.

Table 4: The mean of the minus log predictive densities in the setting where the parameter λ is 1, 10, 50, and 100 and the sample sizes N and M are 50 and 250, respectively. The lowest value in each row is underlined.

λ	AIC	PIC	MSPIC	MSPIC_{BS}
1	−176.77	−201.92	−202.07	<u>−205.40</u>
10	−126.97	<u>−211.60</u>	22.55	−209.33
50	1176.34	−180.16	544.08	<u>−188.78</u>
100	5496.64	−75.54	750.14	<u>−180.80</u>
150	14922.99	−75.41	871.94	<u>−178.16</u>
200	33812.08	38.62	957.92	<u>−182.71</u>

Table 5: The mean of the minus log predictive densities in the setting where the parameter λ is 1, 10, 50, and 100 and the sample sizes N and M are 100 and 500, respectively. The lowest value in each row is underlined.

λ	AIC	PIC	MSPIC	MSPIC_{BS}
1	−418.87	<u>−438.15</u>	<u>−438.15</u>	−436.18
10	−361.78	−418.92	<u>−419.22</u>	−416.92
50	124.53	−408.42	−408.42	<u>−420.98</u>
100	2273.35	−340.89	1287.12	<u>−405.19</u>
150	4437.98	−285.04	1528.31	<u>−392.05</u>
200	9491.38	−191.91	1698.95	<u>−406.93</u>

We compare the minus log plug-in predictive distribution given by

$$\begin{aligned} -\log q_m(y^{(M)}|\hat{\theta}_m(x^{(N)})) &= \frac{M}{2}\log(2\pi) + \frac{M}{2}\log(|x^{(N)\top} - \Phi_m(\Phi_m^\top\Phi_m)^{-1}\Phi_m^\top x^{(N)\top}|^2/N) \\ &\quad + \frac{1}{2}\frac{|y^{(M)\top} - \tilde{\Phi}_m(\Phi_m^\top\Phi_m)^{-1}\Phi_m^\top x^{(N)}|^2}{|x^{(N)\top} - \Phi_m(\Phi_m^\top\Phi_m)^{-1}\Phi_m^\top x^{(N)}|^2/N} \end{aligned}$$

of the AIC-best model and the minus log Bayesian predictive distribution with $\pi(\theta_m, \sigma) = 1/\sigma$ given by

$$\begin{aligned} -\log q_{m,\pi}(y^{(M)}|x^{(N)}) &= \frac{N+M-d_m}{2}\log\left(\left|\begin{pmatrix} x^{(N)\top} \\ y^{(M)\top}\end{pmatrix} - \begin{pmatrix}\Phi_m \\ \tilde{\Phi}_m\end{pmatrix}\hat{\theta}_m(x^{(N)}, y^{(M)})\right|^2\right) \\ &\quad - \frac{N-d_m}{2}\log\left(\left|x^{(N)\top} - \Phi_m\hat{\theta}_m(x^{(N)})\right|^2\right) \\ &\quad + \frac{1}{2}\log\frac{|\Phi_m^\top\Phi_m + \tilde{\Phi}_m^\top\tilde{\Phi}_m|}{|\Phi_m^\top\Phi_m|} - \log\frac{\Gamma(\frac{M+N-d_m}{2})}{\Gamma(\frac{N-d_m}{2})} \end{aligned}$$

of the PIC-best, the MSPIC-best, and the MSPIC_{BS}-best models. The choice of the prior distribution is asymptotically irrelevant according to Theorem 3.1. The reason why we use the above Bayesian distribution is because it is mini-max under the Kullback–Leibler risk. See Liang and Barron (2004). Tables 4 and 5 show that MSPIC_{BS} has the lowest value of the minus log predictive distribution, except for the setting where λ is 10. The dominance of MSPIC_{BS} is enlarged depending on the degree of the extrapolation, i.e., the value of λ.

6 Discussion and Conclusion

In this paper, we have considered the multistep ahead prediction under local misspecification. We have shown that the Bayesian predictive distribution has smaller Kullback–Leibler risk in the setting than the plug-in predictive distribution, regardless of the prior choice. From the

results, we have proposed the information criterion MSPIC for the multistep ahead prediction. The proposed information criterion MSPIC is an asymptotically unbiased estimator of the Kullback–Leibler risk of the Bayesian predictive distribution. By considering the variance of the information criterion MSPIC, we have proposed the bootstrap adjustment $\mathrm{MSPIC}_{\mathrm{BS}}$. Numerical experiments show that our proposed information criterion is effective.

Appendix

In this appendix, we provide proofs of Theorems 3.1 and 4.1. The proofs consist of three parts: the connection formula of the best approximating points (Lemma Appendix.1), the expansions of the maximum likelihood estimators (Lemma Appendix.2), and the expansions of the Kullback–Leibler risk $R(\omega^*, q_{m,\pi})$.

We need some additional notations for the proofs. In the appendix, we write θ instead of θ_m because we fix the submodel $\mathcal{M}_m$ and make expansions easier to see. The simultaneous distribution of $(x^{(N)}, y^{(M)})$ is denoted by $r(x^{(N)}, y^{(M)}|\omega^*)$. In our setting, distribution $r(x^{(N)}, y^{(M)}|\omega^*)$ is given as the product $p(x^{(N)}|\omega^*)q(y^{(M)}|\omega^*)$. We use notations $g^{(r)}(\omega)$ and $g^{(r)}(\theta)$ for the Fisher information matrices of $r(x^{(N)}, y^{(M)}|\omega)$ and $r(x^{(N)}, y^{(M)}|\omega(\theta, 0))$, respectively. Note that $g^{(r)}(\omega) = g^{(p)}(\omega) + g^{(q)}(\omega)$. We denote $g^{(p)}_{a\alpha}\frac{\partial\xi^\alpha}{\partial\omega^s}$ by $g^{(p)}_{as}$ and use $g^{(r)}_{as}$ and $g^{(q)}_{as}$ in the same manner.

We denote the maximum likelihood estimator of $r(x^{(N)}, y^{(M)}|\omega)$ by $\hat{\omega}(x^{(N)}, y^{(M)})$ and the restricted maximum likelihood estimator of $r(x^{(N)}, y^{(M)}|\omega(\theta, 0))$ by $\hat{\theta}(x^{(N)}, y^{(M)})$. We denote embeddings of $\hat{\theta}(x^{(N)})$ and $\hat{\theta}(x^{(N)}, y^{(M)})$ into parameter ω by $\hat{\omega}_m(x^{(N)})$ and $\hat{\omega}_m(x^{(N)}, y^{(M)})$, respectively. We denote the best approximating point of ω^* with respect to $r(x^{(N)}, y^{(M)}|\omega(\theta, 0))$ by $\theta^{(r)}$. In other words, $\theta^{(r)}$ is defined by

$$\theta^{(r)} = \operatorname*{argmax}_{\theta\in\Theta_m} \mathrm{E}_{\omega^*}[\log r(x^{(N)}, y^{(M)}|\omega(\theta, 0))]. \tag{12}$$

In the appendix, we write $\omega^{(p)}$ and $\omega^{(r)}$ instead of $\omega(\theta^{(p)}, 0)$ and $\omega(\theta^{(r)}, 0)$, respectively. We write $\xi^{(p)}$ instead of $\xi(\theta^{(p)}, 0)$.

We denote the (a, b)-components of the observed Fisher information matrices of $p(x^{(N)}|\omega(\theta, 0))$ and $r(x^{(N)}, y^{(M)}|\omega(\theta, 0))$ by $\hat{G}^{(p)}_{ab}(\hat{\theta}(x^{(N)}))$ and $\hat{G}^{(r)}_{ab}(\hat{\theta}(x^{(N)}, y^{(M)}))$, respectively. We denote the stochastic large and small orders with respect to the distribution with the parameter ω by O_ω and o_ω, respectively.

Lemma Appendix.1. *Under local misspecification, the following two equations hold: for* $a \in \{1, \ldots, d_m\}$

$$h^a = -g^{(p)ab}_m(\theta^{(p)})g^{(p)}_{b\kappa}(\xi^{(p)})h^\kappa + \mathrm{O}(1/\sqrt{N}) \tag{13}$$

and

$$\theta^{(r)a}_m - \theta^{(p)a}_m = g^{(r)ab}_m(\theta^{(p)})g^{(r)}_{bs}(\omega^{(p)})\frac{h^s}{\sqrt{N}} + \mathrm{O}(1/N). \tag{14}$$

Proof. First, we show that the former equation holds. From (3), we obtain for $i \in \{1, \ldots, N\}$,

$$p(x_i|\omega^*) = p(x_i|\omega^{(p)})\left[1 + \partial_s \log p(x_i|\omega^{(p)})\frac{h^s}{\sqrt{N}} + \mathrm{O}_{\omega^{(p)}}(1/N)\right], \tag{15}$$

and for $j \in \{1, \ldots, M\}$,

$$q(y_j|\omega^*) = q(y_j|\omega^{(p)})\left[1 + \partial_s \log q(y_j|\omega^{(p)})\frac{h^s}{\sqrt{N}} + \mathrm{O}_{\omega^{(p)}}(1/N)\right], \tag{16}$$

respectively.

Consider the definition of $\omega^{(p)}$:

$$\frac{1}{\sqrt{N}}\mathrm{E}_{\omega^*}\left[\partial_a \log p(x^{(N)}|\omega^{(p)})\right] = 0. \tag{17}$$

From the independence of $x^{(N)}$ and from (15), the LHS in (17) is expanded as

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\mathrm{E}_{\omega^*}\left[\partial_a \log p(x^{(N)}|\omega^{(p)})\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{E}_{\omega^*}\left[\partial_a \log p(x_i|\omega^{(p)})\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{E}_{\omega^{(p)}}\left[\left\{1+\partial_s \log p(x_i|\omega^{(p)})\frac{h^s}{\sqrt{N}}+\mathrm{O}_{\omega^{(p)}}(1/N)\right\}\{\partial_a \log p(x_i|\omega^{(p)})\}\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{E}_{\omega^{(p)}}\left[\partial_a \log p(x_i|\omega^{(p)})\right] \\
&\quad + \sum_{i=1}^{N}\mathrm{E}_{\omega^{(p)}}\left[\partial_s \log p(x_i|\omega^{(p)})\partial_a \log p(x_i|\omega^{(p)})\right]\frac{h^s}{N}+\mathrm{O}(1/\sqrt{N}) \\
&= \frac{1}{N}g_{as}^{(p)}(\omega^{(p)})h^s+\mathrm{O}(1/\sqrt{N}).
\end{aligned} \tag{18}$$

By comparing (17) with (18) up to constant order, we obtain

$$\frac{1}{N}g_{as}^{(p)}(\omega^{(p)})h^s = \mathrm{O}(1/\sqrt{N}).$$

By the reparameterization of ω to ξ, we obtain

$$\frac{1}{N}g_{a\alpha}^{(p)}(\xi^{(p)})h^\alpha = \mathrm{O}(1/\sqrt{N}). \tag{19}$$

Thus we obtain (13).

Next, we show the latter equation holds. Consider the definition of $\omega^{(r)}$:

$$\frac{1}{\sqrt{N}}\mathrm{E}_{\omega^*}\left[\partial_a \log r(x^{(N)},y^{(M)}|\omega^{(r)})\right] = 0. \tag{20}$$

From the independence of $x^{(N)}$ and $y^{(M)}$, from (15) and (16), and from the Taylor expansions of $\partial_a \log p(x_i|\omega^{(r)})$ and $\partial_a \log q(y_j|\omega^{(r)})$ around $\omega^{(p)}$, the LHS in (20) is expanded as

$$\begin{aligned}
&\frac{1}{\sqrt{N}}\mathrm{E}_{\omega^*}\left[\partial_a \log r(x^{(N)},y^{(M)}|\omega^{(r)})\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{E}_{\omega^*}\left[\partial_a \log p(x_i|\omega^{(r)})\right]+\frac{1}{\sqrt{N}}\sum_{j=1}^{M}\mathrm{E}_{\omega^*}\left[\partial_a \log q(y_j|\omega^{(r)})\right] \\
&= \frac{1}{\sqrt{N}}\Sigma_{i=1}^{N}\mathrm{E}_{\omega^{(p)}}\left[\left\{1+\partial_s \log p(x_i|\omega^{(p)})\frac{h^s}{\sqrt{N}}+\mathrm{O}_{\omega^{(p)}}(1/N)\right\}\partial_a \log p(x_i|\omega^{(r)})\right] \\
&\quad +\frac{1}{\sqrt{N}}\Sigma_{j=1}^{M}\mathrm{E}_{\omega^{(p)}}\left[\left\{1+\partial_s \log q(y_j|\omega^{(p)})\frac{h^s}{\sqrt{N}}+\mathrm{O}_{\omega^{(p)}}(1/N)\right\}\partial_a \log q(y_j|\omega^{(r)})\right] \\
&= \frac{1}{\sqrt{N}}\sum_{i=1}^{N}\mathrm{E}_{\omega^{(p)}}\left[\left\{1+\partial_s \log p(x_i|\omega^{(p)})\frac{h^s}{\sqrt{N}}\right\}\right. \\
&\qquad \left.\times\left\{\partial_a \log p(x_i|\omega^{(p)})+\partial_{ab}\log p(x_i|\omega^{(p)})(\theta^{(r)b}-\theta^{(p)b})+\mathrm{O}_{\omega^{(p)}}(||\theta^{(r)}-\theta^{(p)}||^2)\right\}\right] \\
&\quad +\frac{1}{\sqrt{N}}\sum_{j=1}^{M}\mathrm{E}_{\omega^{(p)}}\left[\left\{1+\partial_s \log q(y_j|\omega^{(p)})\frac{h^s}{\sqrt{N}}\right\}\right. \\
&\qquad \left.\times\left\{\partial_a \log q(y_j|\omega^{(p)})+\partial_{ab}\log q(y_j|\omega^{(p)})(\theta^{(r)b}-\theta^{(p)b})+\mathrm{O}_{\omega^{(p)}}(||\theta^{(r)}-\theta^{(p)}||^2)\right\}\right] \\
&= \frac{1}{\sqrt{N}}\mathrm{E}_{\omega^{(p)}}\left[\partial_a \log r(x^{(N)},y^{(M)}|\omega^{(p)})\right] \\
&\quad +g_{sa}^{(r)}(\omega^{(p)})\frac{h^s}{N}-\frac{1}{\sqrt{N}}g_{ab}^{(r)}(\theta^{(p)})(\theta^{(r)b}-\theta^{(p)b})+\mathrm{O}(\sqrt{N}||\theta^{(r)}-\theta^{(p)}||^2).
\end{aligned} \tag{21}$$

Thus, we obtain (14) by comparing (20) with (21) up to constant order. □

Lemma Appendix.2. *Under local misspecification, the following equations hold: for* $a \in \{1, \ldots, d_m\}$,

$$\hat{\theta}^a(x^{(N)}, y^{(M)}) - \theta^{(r)a} = g_m^{(r)ab}(\theta^{(r)})\partial_b \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) + \mathrm{O}_{\omega^*}(1/N) \tag{22}$$

and

$$\hat{\theta}^a(x^{(N)}) - \theta^{(p)a} = g_m^{(p)ab}(\theta^{(p)})\partial_b \log p(x^{(N)}|\omega^{(p)}) + \mathrm{O}_{\omega^*}(1/N), \tag{23}$$

respectively.

For $s \in \{1, \ldots, d_{\mathrm{full}}\}$,

$$\hat{\omega}^s(x^{(N)}, y^{(M)}) - \omega^{*s} = g^{(r)st}(\omega^*)\partial_t \log r(x^{(N)}, y^{(M)}|\omega^*) + \mathrm{O}_{\omega^*}(1/N) \tag{24}$$

and

$$\hat{\omega}^s(x^{(N)}) - \omega^{*s} = g^{(p)st}(\omega^*)\partial_t \log p(x^{(N)}|\omega^*) + \mathrm{O}_{\omega^*}(1/N), \tag{25}$$

respectively.

Proof. Consider the estimative equations:

$$\partial_a \log r(x^{(N)}, y^{(M)}|\hat{\omega}_m(x^{(N)}, y^{(M)})) = 0 \tag{26}$$

and

$$\partial_a \log p(x^{(N)}|\hat{\omega}_m(x^{(N)})) = 0. \tag{27}$$

We apply the Taylor expansions around $\omega^{(r)}$ and $\omega^{(p)}$ to equations (26) and (27), respectively. Since $\partial_{ab} \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) + g_{ab}^{(r)}(\theta^{(r)}) = \mathrm{O}_{\omega^{(r)}}(\sqrt{N})$ and $\omega^* - \omega^{(r)} = \mathrm{O}(1/\sqrt{N})$, we obtain the following expansion:

$$\begin{aligned}
&\partial_a \log r(x^{(N)}, y^{(M)}|\hat{\omega}_m(x^{(N)}, y^{(M)})) \\
&= \partial_a \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) + \partial_{ab} \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) \left\{\hat{\theta}^b(x^{(N)}, y^{(M)}) - \theta^{(r)b}\right\} \\
&\quad + \mathrm{O}_{\omega^{(r)}}(\sqrt{N}||\hat{\theta}(x^{(N)}, y^{(M)}) - \theta^{(r)}||) + \mathrm{O}_{\omega^{(r)}}(N||\hat{\theta}(x^{(N)}, y^{(M)}) - \theta^{(r)}||^2) \\
&= \partial_a \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) - g_{ab}^{(r)}(\theta^{(r)}) \left\{\hat{\theta}^b(x^{(N)}, y^{(M)}) - \theta^{(r)b}\right\} + \mathrm{O}_{\omega^*}(1).
\end{aligned}$$

Likewise, we obtain the following expansion:

$$\begin{aligned}
&\partial_a \log p(x^{(N)}|\hat{\omega}_m(x^{(N)})) \\
&= \partial_a \log p(x^{(N)}|\omega^{(p)}) + \partial_{ab} \log p(x^{(N)}|\omega^{(p)}) \left\{\hat{\theta}^b(x^{(N)}) - \theta^{(p)b}\right\} \\
&\quad + \mathrm{O}_{\omega^{(p)}}(\sqrt{N}||\hat{\theta}(x^{(N)}) - \theta^{(p)}||) + \mathrm{O}_{\omega^{(p)}}(N||\hat{\theta}(x^{(N)}) - \theta^{(p)}||^2) \\
&= \partial_a \log p(x^{(N)}|\omega^{(p)}) - g_{ab}^{(p)}(\theta^{(p)}) \left\{\hat{\theta}^b(x^{(N)}) - \theta^{(p)b}\right\} + \mathrm{O}_{\omega^*}(1).
\end{aligned}$$

Thus, we obtain (22) and (23). Equations (24) and (25) immediately follow from the estimative equations of $\hat{\omega}$. For example, see Theorem 5.39 in van der Vaart (1998). □

Proof of Theorem 3.1. We prove Theorem 3.1 by using the above lemmas. Consider the following decomposition of the Kullback–Leibler risk:

$$R(\omega^*, q_{m,\pi}) = \mathrm{E}_{\omega^*}\left[\log \frac{r(x^{(N)}, y^{(M)}|\omega^*)}{r_{m,\pi}(x^{(N)}, y^{(M)})}\right] - \mathrm{E}_{\omega^*}\left[\log \frac{p(x^{(N)}|\omega^*)}{p_{m,\pi}(x^{(N)})}\right]. \tag{28}$$

The marginal distributions $r_{m,\pi}(x^{(N)}, y^{(M)})$ and $p_{m,\pi}(x^{(N)})$ are expanded as

$$\begin{aligned} r_{m,\pi}(x^{(N)}, y^{(M)}) &= (2\pi)^{d_m/2} \frac{\pi(\hat{\theta}(x^{(N)}, y^{(M)}))}{|\hat{G}^{(r)}(\hat{\theta}(x^{(N)}, y^{(M)}))|^{1/2}} \\ &\quad \times r(x^{(N)}, y^{(M)}|\hat{\omega}_m(x^{(N)}, y^{(M)}))\{1 + \mathrm{o}(1)\} \end{aligned} \tag{29}$$

and

$$p_{m,\pi}(x^{(N)}) = (2\pi)^{d_m/2} \frac{\pi(\hat{\theta}(x^{(N)}))}{|\hat{G}^{(p)}(\hat{\theta}(x^{(N)}))|^{1/2}} p(x^{(N)}|\hat{\omega}_m(x^{(N)}))\{1 + \mathrm{o}(1)\}, \tag{30}$$

respectively. See p. 117 in Ghosh et al. (2006).

By using the marginal expansions (29) and (30), the above decomposition is expanded as

$$\begin{aligned} & R(\omega^*, q_{m,\pi}) \\ &= \mathrm{E}_{\omega^*}\left[\log \frac{r(x^{(N)}, y^{(M)}|\omega^*)}{r(x^{(N)}, y^{(M)}|\hat{\omega}_m(x^{(N)}, y^{(M)}))}\right] - \mathrm{E}_{\omega^*}\left[\log \frac{p(x^{(N)}|\omega^*)}{p(x^{(N)}|\hat{\omega}_m(x^{(N)}))}\right] \\ &\quad + \mathrm{E}_{\omega^*}\left[\frac{1}{2}\log \frac{|\hat{G}^{(r)}(\hat{\theta}(x^{(N)}, y^{(M)}))|}{|\hat{G}^{(p)}(\hat{\theta}(x^{(N)}))|}\right] - \mathrm{E}_{\omega^*}\left[\log \frac{\pi(\hat{\theta}(x^{(N)}, y^{(M)}))}{\pi(\hat{\theta}(x^{(N)}))}\right] + \mathrm{o}(1). \end{aligned} \tag{31}$$

From (3), (14), and (22), the following equation holds:

$$\begin{aligned} & \hat{\omega}_m^s(x^{(N)}, y^{(M)}) - \omega^{*s} \\ &= \omega_m^s(x^{(N)}, y^{(M)}) - \omega^{(r)s} + \omega^{(r)s} - \omega^{(p)s} + \omega^{(p)s} - \omega^{*s} \\ &= \frac{\partial \omega^s}{\partial \theta^a}(\theta^{(r)}) g_m^{(r)ab}(\theta^{(r)}) \partial_b \log r(x^{(N)}, y^{(M)}|\omega^{(r)}) \\ &\quad + \frac{\partial \omega^s}{\partial \theta^a}(\theta^{(p)}) g_m^{(r)ab}(\theta^{(p)}) g_{bs}^{(r)}(\omega^{(p)}) \frac{h^s}{\sqrt{N}} - \frac{h^s}{\sqrt{N}} + \mathrm{O}_{\omega^*}(1/N). \end{aligned} \tag{32}$$

First, consider the first term in (31). By using the Taylor expansion, we expand the negative of the first term as

$$\begin{aligned} & \mathrm{E}_{\omega^*}\left[\log \frac{r(x^{(N)}, y^{(M)}|\hat{\omega}_m(x^{(N)}, y^{(M)}))}{r(x^{(N)}, y^{(M)}|\omega^*)}\right] \\ &= \mathrm{E}_{\omega^*}\left[\partial_s \log r(x^{(N)}, y^{(M)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)}, y^{(M)}) - \omega^{*s}\}\right] \\ &\quad + \frac{1}{2}\mathrm{E}_{\omega^*}\left[\partial_{st} \log r(x^{(N)}, y^{(M)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)}, y^{(M)}) - \omega^{*s}\}\{\hat{\omega}_m^t(x^{(N)}, y^{(M)}) - \omega^{*t}\}\right] \\ &\quad + \mathrm{o}(1). \end{aligned} \tag{33}$$

From the Taylor expansion of $\partial_b \log r(x^{(N)}, y^{(M)}|\omega^{(r)})$ around ω^*, we obtain the following equation for the first term in (33):

$$\begin{aligned} & \mathrm{E}_{\omega^*}[\partial_s \log r(x^{(N)}, y^{(M)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)}, y^{(M)}) - \omega^{*s}\}] \\ &= \mathrm{E}_{\omega^*}\left[\partial_s \log r(x^{(N)}, y^{(M)}|\omega^*)\left\{\frac{\partial \omega^s}{\partial \theta^a}(\theta^{(r)}) g_m^{(r)ab}(\theta^{(r)}) \partial_b \log r(x^{(N)}, y^{(M)}|\omega^{(r)})\right.\right. \\ &\qquad \left.\left. + \frac{\partial \omega^s}{\partial \theta^a}(\theta^{(p)}) g_m^{(r)ab}(\theta^{(p)}) g_{bs}^{(r)}(\omega^{(p)}) \frac{h^s}{\sqrt{N}} - \frac{h^s}{\sqrt{N}}\right\}\right] + \mathrm{o}(1) \\ &= d_m + \mathrm{o}(1). \end{aligned} \tag{34}$$

From (32), we expand the second term in (33) as

$$
\begin{aligned}
&\frac{1}{2}\mathrm{E}_{\omega^*}[\partial_{st}\log r(x^{(N)},y^{(M)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)},y^{(M)})-\omega^{*s}\}\{\hat{\omega}_m^t(x^{(N)},y^{(M)})-\omega^{*t}\}]\\
=\;& -\frac{1}{2}g_{st}^{(r)}(\omega^*)\mathrm{E}_{\omega^*}[\{\hat{\omega}_m^s(x^{(N)},y^{(M)})-\omega^{*s}\}\{\hat{\omega}_m^t(x^{(N)},y^{(M)})-\omega^{*t}\}]+\mathrm{o}(1)\\
=\;& -\frac{1}{2}g_{st}^{(r)}(\omega^*)\mathrm{E}_{\omega^*}\left[\left\{\frac{\partial\omega^s}{\partial\theta^a}(\theta^{(r)})g_m^{(r)ab}(\theta^{(r)})\partial_b\log r(x^{(N)},y^{(M)}|\omega^{(r)})\right.\right.\\
&\qquad\left.+\frac{\partial\omega^s}{\partial\theta^a}(\theta^{(p)})g_m^{(r)ab}(\theta^{(p)})g_{bs}^{(r)}(\omega^{(p)})\frac{h^s}{\sqrt{N}}-\frac{h^s}{\sqrt{N}}\right\}\\
&\qquad\times\left\{\frac{\partial\omega^t}{\partial\theta^c}(\theta^{(r)})g_m^{(r)cd}(\theta^{(r)})\partial_d\log r(x^{(N)},y^{(M)}|\omega^{(r)})\right.\\
&\qquad\left.\left.+\frac{\partial\omega^t}{\partial\theta^c}(\theta^{(p)})g_m^{(r)cd}(\theta^{(p)})g_{dt}^{(r)}(\omega^{(p)})\frac{h^t}{\sqrt{N}}-\frac{h^t}{\sqrt{N}}\right\}\right]+\mathrm{o}(1)\\
=\;& -\frac{1}{2}g_{ac}^{(r)}(\omega^{(r)})g_m^{(r)ab}(\theta^{(r)})g_m^{(r)cd}(\theta^{(r)})\mathrm{E}_{\omega^*}\left[\partial_b\log r(x^{(N)},y^{(M)}|\omega^{(r)})\partial_d\log r(x^{(N)},y^{(M)}|\omega^{(r)})\right]\\
&-\frac{1}{2}g_{st}^{(r)}(\omega^*)\frac{h^sh^t}{N}\\
&-\frac{1}{2}g_{st}^{(r)}(\omega^*)\frac{\partial\omega^s}{\partial\theta^a}(\theta^{(p)})\frac{\partial\omega^t}{\partial\theta^c}(\theta^{(p)})g_m^{(r)ab}(\theta^{(p)})g_{bu}^{(r)}(\omega^{(p)})\frac{h^u}{\sqrt{N}}g_m^{(r)cd}(\theta^{(p)})g_{dv}^{(r)}(\omega^{(p)})\frac{h^v}{\sqrt{N}}\\
&+g_{st}^{(r)}(\omega^*)\frac{h^s}{\sqrt{N}}\frac{\partial\omega^t}{\partial\theta^c}(\theta^{(p)})g_m^{(r)cd}(\theta^{(p)})g_{dv}^{(r)}(\omega^{(p)})\frac{h^v}{\sqrt{N}}+\mathrm{o}(1). \qquad (35)
\end{aligned}
$$

From the independence of $x^{(N)}$ and $y^{(M)}$ and from the Taylor expansions of $\partial_a\log p(x_i|\omega^{(r)})$ and $\partial_a\log q(y_j|\omega^{(r)})$ around ω^*,

$$
\begin{aligned}
&\mathrm{E}_{\omega^*}\left[\partial_a\log r(x^{(N)},y^{(M)}|\omega^{(r)})\partial_b\log r(x^{(N)},y^{(M)}|\omega^{(r)})\right]\\
=\;& \Sigma_{i=1}^N\mathrm{E}_{\omega^*}[\partial_a\log p(x_i|\omega^{(r)})\partial_b\log p(x_i|\omega^{(r)})]+\Sigma_{j=1}^M\mathrm{E}_{\omega^*}[\partial_a\log q(y_j|\omega^{(r)})\partial_b\log q(y_j|\omega^{(r)})]\\
&+\Sigma_{i\neq k}^N\mathrm{E}_{\omega^*}[\partial_a\log p(x_i|\omega^{(r)})\partial_b\log p(x_k|\omega^{(r)})]+\Sigma_{j\neq l}^M\mathrm{E}_{\omega^*}[\partial_a\log q(y_j|\omega^{(r)})\partial_b\log q(y_l|\omega^{(r)})]\\
&+2\Sigma_{i,j}^{i=N,j=M}\mathrm{E}_{\omega^*}[\partial_a\log p(x_i|\omega^{(r)})\partial_b\log q(y_j|\omega^{(r)})]\\
=\;& g_{ab}^{(r)}(\omega^{(r)})+\mathrm{O}(\sqrt{N}). \qquad (36)
\end{aligned}
$$

By substituting (36) into the first term in (35), we obtain the following further expansion of (35):

$$
\begin{aligned}
&\frac{1}{2}\mathrm{E}_{\omega^*}[\partial_{st}\log r(x^{(N)},y^{(M)}|\omega^*)\{\omega^{*s}-\hat{\omega}_m^s(x^{(N)},y^{(M)})\}\{\omega^{*t}-\hat{\omega}_m^t(x^{(N)},y^{(M)})\}]\\
=\;& -\frac{1}{2}g_{st}^{(r)}(\omega^{(p)})\frac{h^sh^t}{N}+\frac{1}{2}g_m^{(r)ab}(\theta^{(p)})g_{as}^{(r)}(\omega^{(p)})g_{bt}^{(r)}(\omega^{(p)})\frac{h^sh^t}{N}-\frac{1}{2}d_m+\mathrm{o}(1). \qquad (37)
\end{aligned}
$$

By combining (34) and (37), we obtain the following equation for (33):

$$
\begin{aligned}
&\mathrm{E}_{\omega^*}\left[\log\frac{r(x^{(N)},y^{(M)}|\hat{\omega}_m(x^{(N)},y^{(M)}))}{r(x^{(N)},y^{(M)}|\omega^*)}\right]\\
=\;& -\frac{1}{2}\left[g_{st}^{(r)}(\omega^{(p)})-g_m^{(r)ab}(\theta^{(p)})g_{as}^{(r)}(\omega^{(p)})g_{bt}^{(r)}(\omega^{(p)})\right]\frac{h^sh^t}{N}+\frac{1}{2}d_m+\mathrm{o}(1). \qquad (38)
\end{aligned}
$$

Next, consider the second term in (31). The estimator $\hat{\omega}_m(x^{(N)})$ is expanded as

$$
\hat{\omega}_m^s(x^{(N)})-\omega^{*s} = \frac{\partial\omega^s}{\partial\theta^a}(\theta^{(p)})g_m^{(p)ab}(\theta^{(p)})\partial_b\log p(x^{(N)}|\omega^{(p)})-\frac{h^s}{\sqrt{N}}+\mathrm{O}_{\omega^*}(1/N). \qquad (39)
$$

By using the Taylor expansion, we expand the negative of the second term in (31) as

$$\begin{aligned}
&\mathrm{E}_{\omega^*}\left[\log\frac{p(x^N|\hat{\omega}_m(x^{(N)}))}{p(x^{(N)}|\omega^*)}\right]\\
=\;&\mathrm{E}_{\omega^*}[\partial_s \log p(x^{(N)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)})-\omega^{*s}\}]\\
&+\frac{1}{2}\mathrm{E}_{\omega^*}[\partial_{st}\log p(x^{(N)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)})-\omega^{*s}\}\{\hat{\omega}_m^t(x^{(N)})-\omega^{*t}\}]+\mathrm{o}(1). \qquad (40)
\end{aligned}$$

From (39), we obtain

$$\begin{aligned}
&\mathrm{E}_{\omega^*}\left[\partial_s \log p(x^{(N)}|\omega^*)\{\hat{\omega}_m^s(x^{(N)})-\omega^{*s}\}\right]\\
=\;&\mathrm{E}_{\omega^*}\left[\partial_s \log p(x^{(N)}|\omega^*)\left\{\frac{\partial\omega^s}{\partial\theta^a}(\theta^{(p)})g_m^{(p)ab}(\theta^{(p)})\partial_b \log p(x^{(N)}|\omega^{(p)})-\frac{h^s}{\sqrt{N}}+\mathrm{O}_{\omega^*}(1/N)\right\}\right]\\
=\;&g_{ab}^{(p)}(\omega^{(p)})g_m^{(p)ab}(\theta^{(p)})+\mathrm{o}(1)\\
=\;&d_m+\mathrm{o}(1) \qquad (41)
\end{aligned}$$

and

$$\begin{aligned}
&\mathrm{E}_{\omega^*}[g_{st}^{(p)}(\omega^*)\{\hat{\omega}_m(x^{(N)})-\omega^{*s}\}\{\hat{\omega}_m(x^{(N)})-\omega^{*t}\}]\\
=\;&g_{st}^{(p)}(\omega^*)\frac{h^s h^t}{N}+g_{st}^{(p)}(\omega^*)\frac{\partial\omega^s}{\partial\theta^a}(\theta^{(p)})\frac{\partial\omega^t}{\partial\theta^b}(\theta^{(p)})g_m^{(p)ac}(\theta^{(p)})g_m^{(p)bd}(\theta^{(p)})g_{bd}^{(p)}(\omega^{(p)})+\mathrm{o}(1)\\
=\;&g_{st}^{(p)}(\omega^*)\frac{h^s h^t}{N}+d_m+\mathrm{o}(1). \qquad (42)
\end{aligned}$$

From (41) and (42), we obtain the following equation for (40):

$$\mathrm{E}_{\omega^*}\left[\log\frac{p(x^{(N)}|\hat{\omega}_m(x^{(N)}))}{p(x^{(N)}|\omega^*)}\right] \;=\; -\frac{1}{2}g_{st}^{(p)}(\omega^*)\frac{h^s h^t}{N}+\frac{1}{2}d_m+\mathrm{o}(1). \qquad (43)$$

The Taylor expansions around $\theta^{(p)}$ and equation (14) show that the third and fourth terms in (31) are equal to o(1). Thus, from (38) and (43), the Kullback–Leibler risk $R(\omega^*, q_{m,\pi})$ is expanded as

$$\begin{aligned}
&R(\omega^*, q_{m,\pi})\\
=\;&\frac{1}{2N}\left[g_{st}^{(r)}(\omega^*)-g_{st}^{(p)}(\omega^*)-g_m^{(r)ab}(\theta^{(p)})g_{sa}^{(r)}(\omega^{(p)})g_{tb}^{(r)}(\omega^{(p)})\right]h^s h^t\\
&+\frac{1}{2}\log\frac{|g^{(r)}(\theta^{(p)})|}{|g^{(p)}(\theta^{(p)})|}+\mathrm{o}(1). \qquad (44)
\end{aligned}$$

Note that this is invariant up to o(1) under the reparameterization of ω.

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

$$P_{\alpha\beta} \;=\; g_{\alpha\beta}^{(r)}(\xi^*)-g_{\alpha\beta}^{(p)}(\xi^*)-g_m^{(r)ab}(\theta^{(p)})g_{a\alpha}^{(r)}(\xi^{(p)})g_{b\beta}^{(r)}(\xi^{(p)}). \qquad (45)$$

To complete the proof of Theorem 3.1, we show

$$P_{\alpha\beta}h^\alpha h^\beta/N \;=\; S_{\alpha\beta}h^\alpha h^\beta/N+\mathrm{o}(1). \qquad (46)$$

From (13), we obtain

$$\begin{aligned}
&P_{ab}h^a h^b\\
=\;&\left\{g_{ab}^{(r)}(\xi^{(p)})-g_{ab}^{(p)}(\xi^{(p)})-g_m^{(r)cd}(\theta^{(p)})g_{ac}^{(r)}(\xi^{(p)})g_{bd}^{(r)}(\xi^{(p)})\right\}h^a h^b\\
=\;&-g_{ab}^{(p)}(\xi^{(p)})h^a h^b\\
=\;&-g_{ab}^{(p)}(\xi^{(p)})\left\{-g_m^{(p)ac}(\theta^{(p)})g_{c\kappa}^{(p)}(\xi^{(p)})h^\kappa+\mathrm{o}(1)\right\}\left\{-g_m^{(p)bd}(\theta^{(p)})g_{d\lambda}^{(p)}(\xi^{(p)})h^\lambda+\mathrm{o}(1)\right\}\\
=\;&-g_m^{(p)ab}(\theta^{(p)})g_{a\kappa}^{(p)}(\xi^{(p)})g_{b\lambda}^{(p)}(\xi^{(p)})h^\kappa h^\lambda+\mathrm{o}(N) \qquad (47)
\end{aligned}$$

and

$$
\begin{aligned}
& P_{a\kappa}h^a h^\kappa \\
&= \left\{ g^{(r)}_{a\kappa}(\xi^{(p)}) - g^{(p)}_{a\kappa}(\xi^{(p)}) - g^{(r)cd}_{m}(\theta^{(p)}) g^{(r)}_{ac}(\xi^{(p)}) g^{(r)}_{d\kappa}(\xi^{(p)}) \right\} h^a h^\kappa \\
&= \left\{ g^{(r)}_{a\kappa}(\xi^{(p)}) - g^{(p)}_{a\kappa}(\xi^{(p)}) - g^{(r)}_{a\kappa}(\xi^{(p)}) \right\} \left\{ -g^{(p)ae}_{m}(\theta^{(p)}) g^{(p)}_{e\lambda}(\xi^{(p)}) h^\lambda + \mathrm{o}(1) \right\} h^\kappa \\
&= g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(p)ab}_{m}(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda + \mathrm{o}(N). \qquad (48)
\end{aligned}
$$

We have

$$
P_{\kappa\lambda}h^\kappa h^\lambda = \left\{ g^{(r)}_{\kappa\lambda}(\xi^{(p)}) - g^{(p)}_{\kappa\lambda}(\xi^{(p)}) - g^{(r)ab}_{m}(\theta^{(p)}) g^{(r)}_{a\kappa}(\xi^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) \right\} h^\kappa h^\lambda. \qquad (49)
$$

From (47), (48), and (49), we obtain

$$
\begin{aligned}
P_{\alpha\beta}h^\alpha h^\beta &= P_{ab}h^a h^b + 2P_{a\kappa}h^a h^\kappa + P_{\kappa\lambda}h^\kappa h^\lambda \\
&= \{ g^{(q)}_{\kappa\lambda}(\xi^{(p)}) + g^{(p)ab}_{m}(\theta^{(p)}) g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) \\
&\qquad - g^{(r)ab}_{m}(\theta^{(p)}) g^{(r)}_{a\kappa}(\xi^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) \} h^\kappa h^\lambda + \mathrm{o}(N). \qquad (50)
\end{aligned}
$$

By applying Sherman–Morisson–Woodbury identity to matrix S, the following equation holds:

$$
\begin{aligned}
S &= \left[g^{(q)-1}(\xi^*) + \begin{pmatrix} g^{(p)-1}_{m}(\theta^{(p)}) & 0^\top_{(d_{\mathrm{full}}-d_m)\times d_m} \\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} & 0_{(d_{\mathrm{full}}-d_m)\times(d_{\mathrm{full}}-d_m)} \end{pmatrix} \right]^{-1} \\
&= \left[g^{(q)-1}(\xi^*) + \begin{pmatrix} I \\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} g^{(p)-1}_{m}(\theta^{(p)}) \begin{pmatrix} I & 0^\top_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} \right]^{-1} \\
&= g^{(q)}(\xi^*) \\
&\quad - g^{(q)}(\xi^*) \begin{pmatrix} I \\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} \left[g^{(p)}_{m}(\theta^{(p)}) + \begin{pmatrix} I & 0^\top_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} g^{(q)}(\xi^*) \begin{pmatrix} I \\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} \right]^{-1} \\
&\qquad \begin{pmatrix} I & 0^\top_{(d_{\mathrm{full}}-d_m)\times d_m} \end{pmatrix} g^{(q)}(\xi^*) \\
&= g^{(q)}(\xi^*) - g^{(q)}(\xi^*) \begin{pmatrix} g^{(r)-1}_{m}(\theta^{(p)}) & 0^\top_{(d_{\mathrm{full}}-d_m)\times d_m} \\ 0_{(d_{\mathrm{full}}-d_m)\times d_m} & 0_{(d_{\mathrm{full}}-d_m)\times(d_{\mathrm{full}}-d_m)} \end{pmatrix} g^{(q)}(\xi^*), \qquad (51)
\end{aligned}
$$

where I is the d_m-dimensional identity matrix, $0_{(d_{\mathrm{full}}-d_m)\times d_m}$ is the $(d_{\mathrm{full}}-d_m)\times d_m$-dimensional zero matrix, and $0_{(d_{\mathrm{full}}-d_m)\times(d_{\mathrm{full}}-d_m)}$ is the $(d_{\mathrm{full}}-d_m)\times(d_{\mathrm{full}}-d_m)$-dimensional zero matrix. From (13), we obtain

$$
\begin{aligned}
& S_{ab}h^a h^b \\
&= g^{(q)}_{ac}(\xi^{(p)}) g^{(r)cd}_{m}(\theta^{(p)}) g^{(r)}_{db}(\xi^{(p)}) h^a h^b - g^{(q)}_{ac}(\xi^{(p)}) g^{(r)cd}_{m}(\theta^{(p)}) g^{(q)}_{db}(\xi^{(p)}) h^a h^b \\
&= g^{(p)}_{ac}(\xi^{(p)}) g^{(r)cd}_{m}(\theta^{(p)}) g^{(q)}_{bd}(\xi^{(p)}) h^a h^b \\
&= g^{(p)}_{ac}(\xi^{(p)}) g^{(r)cd}_{m}(\theta^{(p)}) g^{(q)}_{bd}(\xi^{(p)}) \left\{ -g^{(p)ae}_{m}(\theta^{(p)}) g^{(p)}_{e\kappa}(\xi^{(p)}) h^\kappa + \mathrm{o}(1) \right\} \left\{ -g^{(p)bf}_{m}(\theta^{(p)}) g^{(p)}_{f\lambda}(\xi^{(p)}) h^\lambda + \mathrm{o}(1) \right\} \\
&= g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(r)ab}_{m}(\theta^{(p)}) g^{(q)}_{bc}(\xi^{(p)}) g^{(p)cd}_{m}(\theta^{(p)}) g^{(p)}_{d\lambda}(\xi^{(p)}) h^\kappa h^\lambda + \mathrm{o}(N) \\
&= g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(r)ab}_{m}(\theta^{(p)}) \{ g^{(r)}_{bc}(\xi^{(p)}) - g^{(p)}_{bc}(\xi^{(p)}) \} g^{(p)cd}_{m}(\theta^{(p)}) g^{(p)}_{d\lambda}(\xi^{(p)}) h^\kappa h^\lambda + \mathrm{o}(N) \\
&= g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(p)ab}_{m}(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda - g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(r)ab}_{m}(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda + \mathrm{o}(N). \qquad (52)
\end{aligned}
$$

From (13) and the relationship that $g^{(q)} = g^{(r)} - g^{(p)}$, we have

$$\begin{aligned}
& S_{a\kappa} h^a h^\kappa \\
&= g^{(p)}_{ac}(\xi^{(p)}) g^{(r)cd}_m(\theta^{(p)}) g^{(q)}_{d\kappa}(\xi^{(p)}) h^a h^\kappa \\
&= g^{(p)}_{ac}(\xi^{(p)}) g^{(r)cd}_m(\theta^{(p)}) g^{(q)}_{d\kappa}(\xi^{(p)}) \left\{ -g^{(p)ae}_m(\theta^{(p)}) g^{(p)}_{e\lambda}(\xi^{(p)}) h^\lambda + \mathrm{o}(1) \right\} h^\kappa \\
&= -g^{(p)}_{ac}(\xi^{(p)}) g^{(r)cd}_m(\theta^{(p)}) \left\{ g^{(r)}_{d\kappa}(\xi^{(p)}) - g^{(p)}_{d\kappa}(\xi^{(p)}) \right\} g^{(p)ae}_m(\theta^{(p)}) g^{(p)}_{e\lambda}(\xi^{(p)}) h^\lambda h^\kappa + \mathrm{o}(N) \\
&= -g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda \\
&\quad + g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda + \mathrm{o}(N)
\end{aligned} \tag{53}$$

and

$$\begin{aligned}
& S_{\kappa\lambda} h^\kappa h^\lambda \\
&= \left[g^{(q)}_{\kappa\lambda}(\xi^{(p)}) - g^{(q)}_{\kappa a}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(q)}_{b\lambda}(\xi^{(p)}) \right] h^\kappa h^\lambda \\
&= \left\{ g^{(q)}_{\kappa\lambda}(\xi^{(p)}) - \left\{ g^{(r)}_{\kappa a}(\xi^{(p)}) - g^{(p)}_{\kappa a}(\xi^{(p)}) \right\} g^{(r)ab}_m(\theta^{(p)}) \left\{ g^{(r)}_{b\lambda}(\xi^{(p)}) - g^{(p)}_{b\lambda}(\xi^{(p)}) \right\} \right\} h^\kappa h^\lambda \\
&= g^{(q)}_{\kappa\lambda}(\xi^{(p)}) h^\kappa h^\lambda - g^{(r)}_{\kappa a}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda - g^{(p)}_{\kappa a}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda \\
&\quad + 2 g^{(p)}_{\kappa a}(\xi^{(p)}) g^{(r)ab}_m(\theta^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) h^\kappa h^\lambda .
\end{aligned} \tag{54}$$

From (52), (53), and (54), we obtain the following equation:

$$S_{\alpha\beta} h^\alpha h^\beta = \left\{ g^{(q)}_{\kappa\lambda}(\xi^{(p)}) + g^{(p)}_{a\kappa}(\xi^{(p)}) g^{(p)}_m(\theta^{(p)}) g^{(p)}_{b\lambda}(\xi^{(p)}) - g^{(r)}_{a\kappa}(\xi^{(p)}) g^{(r)}_m(\theta^{(p)}) g^{(r)}_{b\lambda}(\xi^{(p)}) \right\} h^\kappa h^\lambda + \mathrm{o}(N).$$

Thus, we obtain (46) and complete the proof of Theorem 3.1. □

Proof of Theorem 4.1. Since $\hat{h}^\alpha$ is decomposed as

$$\begin{aligned}
\frac{\hat{h}^\alpha}{\sqrt{N}} &= \hat{\xi}^\alpha(x^{(N)}) - \xi^{*\alpha} + \xi^{*\alpha} - \xi^{(p)\alpha} + \xi^{(p)\alpha} - \xi^\alpha(\hat{\theta}(x^{(N)}), 0) \\
&= g^{(p)\alpha\beta}(\xi^*) \partial_\beta \log p(x^{(N)}|\xi^*) + \frac{h^\alpha}{\sqrt{N}} \\
&\quad - \delta^\alpha_a g^{(p)ab}_m(\theta^{(p)}) \partial_b \log p(x^{(N)}|\omega^{(p)}) + \mathrm{O}(1/N),
\end{aligned} \tag{55}$$

the expectation of $\hat{S}_{\alpha\beta} \hat{h}^\alpha \hat{h}^\beta / N$ is given as

$$\begin{aligned}
& \mathrm{E}_{\omega^*}[\hat{S}_{\alpha\beta} \hat{h}^\alpha \hat{h}^\beta]/N \\
&= \mathrm{E}_{\omega^*}[S_{\alpha\beta} \hat{h}^\alpha \hat{h}^\beta]/N + \mathrm{o}(1) \\
&= \mathrm{E}_{\omega^*}\left[S_{\alpha\beta} \left\{ g^{(p)\alpha\gamma}(\xi^*) \partial_\gamma \log p(x^{(N)}|\xi^*) + \frac{h^\alpha}{\sqrt{N}} - \delta^\alpha_a g^{(p)ac}_m(\theta^{(p)}) \partial_c \log p(x^{(N)}|\omega^{(p)}) \right\} \right. \\
&\qquad \left. \times \left\{ g^{(p)\beta\delta}(\xi^*) \partial_\delta \log p(x^{(N)}|\xi^*) + \frac{h^\beta}{\sqrt{N}} - \delta^\beta_b g^{(p)bd}_m(\theta^{(p)}) \partial_d \log p(x^{(N)}|\omega^{(p)}) \right\} \right] \\
&\quad + \mathrm{o}(1) \\
&= S_{\alpha\beta} \frac{h^\alpha h^\beta}{N} + S_{\alpha\beta} g^{(p)\alpha\beta}(\xi^*) + S_{ab} g^{(p)ab}_m(\theta^{(p)}) - 2 S_{ab} g^{(p)ab}_m(\theta^{(p)}) + \mathrm{o}(1) \\
&= S_{\alpha\beta} \frac{h^\alpha h^\beta}{N} + S_{\alpha\beta} g^{(p)\alpha\beta}(\xi^*) - S_{ab} g^{(p)ab}_m(\theta^{(p)}) + \mathrm{o}(1).
\end{aligned}$$

Thus, we complete the proof. □

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov and F. Caski, eds., *Proc. of the 2nd international symposium of information theory.* Akadimiai Kiado, pp. 267–281.

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

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

Ghosh, J. K., Delampady, M., and Samanta, T. (2006). *An Introduction to Bayesian Analysis Theory and Methods.* Springer Science+Business Media, New York.

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

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

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

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

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

Konishi, S. and Kitagawa, G. (2003). Asymptotic theory for information criteria in model selection–functional approach. *J. Statist. Plann. and Infer.* **114**, pp. 45–61.

Leeb, H. and Pötscher, B. M. (2005). Model selection and inference: Facts and fiction. *Econom. Theory* **21**, pp. 21–59.

Liang, F. and Barron, A. (2004). Exact minimax strategies for predictive density estimation, data compression, and model selection. *IEEE TRAN. ON INFOR. THEORY* **50**, pp. 2708–2726.

Lv, L. and Liu, J. (2014). Model selection principles in misspecified models. *J. R. Statrist. Soc. B* **76**, pp. 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**, pp. 2523–2534.

Shibata, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, pp. 45–54.

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

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

van der Vaart (1998). *Asymptotic statistics.* Cambridge University Press, New York.